UNIVERSITY
OF SUSSEX

DOCTORAL THESIS

Observation of the associated production of the Higgs boson with a top quark pair with the ATLAS experiment at the LHC

A study in the boosted regime of the $H \to b\bar{b}$ decay channel

A thesis submitted to the University of Sussex
in fulfilment of the requirements for the degree of
Doctor of Philosophy

in the

Department of Physics and Astronomy
School of Mathematical and Physical Sciences

Author:
Emma WINKELS

Supervisor:
Dr. Lily ASQUITH

2nd September 2019

Opgedragen aan mijn 17-jarige zelf,
die thuiskwam na een bezoek aan CERN en riep:
“Dit is het antwoord op alles!”
en niet opgaf tot ze daar onderzoek kon doen.

ACKNOWLEDGEMENTS

A good relationship with your supervisor is essential for surviving your PhD and I am very grateful for having had Lily to accompany me. She has been a great supervisor over the last few years, always helpful professionally and ready for a pub visit to forget about our struggles. A big thanks goes also to Sarah, who together with Lily got me settled into the analysis team. I am also grateful to Carlos, who guided me in my technical work to become an ATLAS author.

The other PhD students at Sussex have made these four years a lot of fun, and long days spent in the office less tedious. Some good nights out were had with Sam and mini Fab, both in Brighton and at CERN. A special thanks goes to Fabio, who I have perhaps spent the most time with over the last four years. It was great to have a friend to share the office with, and many many beers. Thanks to Diana for being a great friend and for sharing our troubles, both professionally and personally. I'd also like to thank Guillermo and Vangelis, who are at a different university but have made the conferences and schools over the years a lot more interesting.

Brighton would not have been the same without the company of Anthony, Ross, Manuel, and Bernardo, who I've shared lots of experiences with since I met them at a Sussex workshop. Whether or not I learned how to become an effective researcher, I gained some great friends.

My time at CERN was exciting and a period of meeting lots of great people. Andrea was one of the first people I met there, she has been a great friend and introduced me to lots of other CERNies that became friends as well. I remember very fondly the nights at R1 and OB with her, Fabio, Ceci, Roger, Sam, Anne, Jose, Lorenzo, Oksana, Alessandro, Alex C., Alex R., Lena, João, and the rest of the Portuguese mafia.

I'd like to thank Sol, Euan, Bene, and Adam for helping me get through my MSc both physics-wise and sanity-wise, and for the organised and unexpected reunions over the years.

Thanks to my friends back home Merel, Puck, Harie, Thomas, and Feline for coming to visit me in my new homes and brightening up some weekends. Thanks to Ginger and Kathelijn for being amazing people and always available for a talk. After living abroad for a few years it is great to have such friends that make my visits back to Amsterdam always memorable.

A huge thanks goes to my family, for always being there for me. Your faith and confidence in me have provided a safety net without which I could not have ventured this far. Thanks mam and Casper for always providing a place to stay when I need it and being amazing hosts, friends, and family all in one. Thanks mam for lending a shoulder to cry on in difficult times both during the MSc and the PhD. Thanks to pap and Zos for being friends as well as family, for coming to visit during my years abroad, and for being first in line for dinners back in Amsterdam.

Last but certainly not least, I want to thank João, whose love, support, and patience have helped me through the difficult phases of the PhD. He has always been there for me with home-cooked meals during busy times, hugs in between meetings at CERN, and many-hour-long Skype calls during our times apart. I am so happy that those Skype calls can now be replaced by real talks.

STATEMENT

I, Emma WINKELS, hereby declare that this thesis has not been, and will not be, submitted in whole or in part to another university for the award of any other degree.

Brighton,
2nd September 2019

Emma WINKELS

Observation of the associated production of the Higgs boson with a top quark pair with the ATLAS experiment at the LHC

by Emma WINKELS

ABSTRACT

Probing the coupling of the Higgs boson to the heaviest known fermion, the top quark, is crucial for testing the Standard Model (SM) and for constraining new physics models. This thesis presents a search for the $t\bar{t}H$ process which gives direct access to this Yukawa coupling. The analysis is optimised for the Higgs decaying to a pair of bottom quarks and uses 36.1 fb^{-1} of data taken at a centre-of-mass energy of 13 TeV with the ATLAS detector.

One of the main challenges of this analysis is the combinatorial ambiguity from the many jets in the final state which makes it difficult to reconstruct the Higgs boson. The boosted analysis specifically targets final states with high transverse momentum in which the decay products of the Higgs boson and/or hadronically decaying top quark are produced collimated into large jets. This gives access to different kinematics and a simplified combinatorial background.

To select events rich in $t\bar{t}H$, we apply cuts on the number of jets and b-jets (jets tagged as containing b-hadrons). The boosted analysis also requires two large jets in each event which are constructed with the reclustering method. Since the $t\bar{t}$ background is overwhelmingly large compared to the $t\bar{t}H$ signal, multivariate techniques are used to discriminate between signal and background events. A Boosted Decision Tree (BDT) is used with eight variables including the Higgs candidate mass. The boosted analysis is combined with the resolved selection to obtain a ratio of the measured cross-section to the SM expectation of $\mu_{t\bar{t}H} = 0.84^{+0.64}_{-0.61}$. This corresponds to a significance of 1.4σ, with an expectation of 1.6σ. A $t\bar{t}H$ signal strength larger than 2.0 is excluded at the 95% confidence level.

The analysis in the $b\bar{b}$ decay channel is combined with three other $t\bar{t}H$ searches optimised for the multilepton, $\gamma\gamma$, and ZZ decay modes. The combination in which all analyses use 36.1 fb^{-1} of data leads to a significance of 4.2σ. This constitutes evidence for $t\bar{t}H$ production and corresponds to a cross-section of $\sigma_{t\bar{t}H} = 590^{+160}_{-150}$ fb which is compatible with the SM prediction of 507^{+35}_{-50} fb. The combination is repeated with the $\gamma\gamma$ and ZZ decay channels updated to include ATLAS data from 2017 and inclusion of the Run I dataset. This results in an observed (expected) significance of 6.3σ (5.1σ) which marks the first direct observation of the Higgs coupling to the top quark.

CONTENTS

INTRODUCTION

The discovery of the Higgs boson in 2012 by the ATLAS and CMS collaborations was an important milestone for the Standard Model of particle physics. This model was described in the 1960s and 1970s and has since been confirmed by a plethora of experimental measurements. The observation of the Higgs boson was the last cornerstone in this experimental verification.

Ever since its discovery, the Higgs boson's properties have been put to the test with the use of data from the Large Hadron Collider (LHC) at CERN. One of the important goals is to understand how the Higgs boson couples to the other elementary particles. The masses of the fermions are a consequence of their coupling with the Higgs field and these masses are therefore proportional to the coupling strength. Since the top quark is the heaviest particle in the Standard Model, it is predicted to have the strongest coupling to the Higgs boson. The measurement of this coupling is important to test the predictions from the Standard Model theory. Any deviations from the expectations of this measurement could point to an exciting arena of new physics.

The top Yukawa coupling can be measured through different processes. The Higgs production via gluon-gluon fusion is one candidate since it is usually mediated by a top quark-loop. However, the loops in this process can hide interesting Beyond the Standard Model effects and any measurement involving such loops needs to make assumptions on the models involved. Therefore, a direct tree-level access to the top Yukawa coupling is necessary in order to disentangle new physics from Standard Model physics and reduce the model dependence of the measurement. The Higgs boson produced in association with two top quarks, the $t\bar{t}H$ process, provides this access and is the topic of this thesis. At the current centre-of-mass energy of the LHC of 13 TeV, the $t\bar{t}H$ cross-section is two orders of magnitude smaller than the gluon-gluon fusion process; it constitutes about 1% of the total Higgs cross-section. However, this cross-section has increased by a factor of four compared to the maximum Run I energy of 8 TeV. The Run II dataset has opened the door to a previously inaccessible region of physics which includes the possibility of observing $t\bar{t}H$.

The work presented here is focused around the decay of the Higgs boson to two bottom quarks, which has the largest branching ratio in the Standard Model. The top quarks in the $t\bar{t}H$ process can decay to a lepton-neutrino pair (leptonic decay) or to a pair of quarks (hadronic

decay). This thesis gives specific focus to the boosted semileptonic decay channel in which the Higgs boson and the hadronically decaying top quark are produced at high transverse momentum compared to their mass. The boosted regime reduces the sensitivity to one of the main problems of the $t\bar{t}H(H \to b\bar{b})$ analysis: the fact that this process has many jets in its final state. The large number of jets leads to complicated combinatorics when reconstructing the objects in the events. In the boosted regime, many of the final-state jets are combined together into large jets which leads to a simplified combinatorial background which in turn can help to improve the sensitivity of the analysis.

The analysis is carried out using 36.1 fb^{-1} of ATLAS data taken in 2015–2016 at a centre-of-mass energy of 13 TeV. The events are categorised according to their number of leptons, (large) jets, and jets containing b-hadrons. The $t\bar{t}H(H \to b\bar{b})$ channel suffers from large backgrounds, the most challenging of which is the production of top quark pairs with additional jets. In order to separate the signal events from the overwhelming background, multivariate techniques are used. The boosted region relies on a Boosted Decision Tree to get as much information as possible from each event and thereby aid in the signal/background classification. The results presented here are an inclusive measurement of the combined resolved and boosted $t\bar{t}H(H \to b\bar{b})$ channels. A combination with three other ATLAS $t\bar{t}H$ analyses optimised for different Higgs decay modes is also included.

This thesis is structured in nine chapters. The theoretical foundations of the Standard Model are summarised in chapter 1, along with details about the Higgs boson and the top quark. Chapter 2 introduces the LHC and the ATLAS detector with which all data used in this thesis was collected. The generation of Monte Carlo samples for the prediction of the signal and background processes is discussed in the first half of chapter 3. The second half of this chapter describes the software framework used for the reconstruction of physics objects from detector signals. Chapter 4 is dedicated to the definition, reconstruction, and calibration of jets which play a large role in the analysis.

The main analysis of this thesis is presented in chapter 5. This chapter discusses the event selection and categorisation of the $t\bar{t}H(H \to b\bar{b})$ analysis. The optimisation of the boosted channel and the multivariate techniques are discussed in detail. Afterwards, chapter 6 gives an overview of the statistical methods used to obtain the final results. These are acquired from a profile likelihood fit which is discussed in chapter 7 along with all the uncertainties used in the fit. Chapter 8 presents the results of the $H \to b\bar{b}$ decay channel individually and in a combination with three other Higgs decay mode analyses. The final remarks and outlook for the future are summarised in chapter 9.

My contribution includes work done on the boosted single-lepton channel of the $t\bar{t}H(H \to b\bar{b})$ analysis. I have carried out the studies in section 3.1.2, in which the full detector simulation is compared to the fast simulation for the production of simulated Monte Carlo samples. From these studies, it was concluded that we need to use the full simulation for the boosted channel. The studies comparing trimmed large jets with reclustered large jets in section 4.5 are also my work. These studies have led to the switch from trimmed to reclustered jets for the boos-

ted analysis. Since the boosted channel was included in the $t\bar{t}H(H \to b\bar{b})$ analysis for the first time, a signal region had to be designed. I designed a region which uses reclustered jets and was chosen amongst three options to be used as the final analysis region. The details of the signal region definition and optimisation are described in section 5.5. I was responsible for the definition and optimisation of our multivariate analysis, as described in section 5.7.2. During the development of the boosted analysis, I ran the full fit (see section 7.1) at each stage in order to check the boosted-only results and the effect of the boosted analysis on the combined results including the resolved channels. I also included my study on the uncertainties of high transverse momentum jets in section 7.3.1. All plots and figures in this thesis are my own, unless a reference to another source is given.

1 THE HIGGS BOSON IN THE STANDARD MODEL

This chapter presents an introduction to the Higgs boson and the Standard Model (SM). It gives an overview of the theoretical foundations of particle physics relevant to the work presented in this thesis. The particle content and fundamental interactions of the SM are discussed in section 1.1. The Brout-Englert-Higgs mechanism (also known as the Higgs mechanism) by which the massive fermions and bosons acquire their mass is described in section 1.2. Since this thesis is concerned with the study of the $t\bar{t}H$ process, sections 1.3 and 1.4 give more details about the Higgs boson and top quark, respectively. The coupling between these two fundamental particles is discussed in section 1.5. All units in this chapter are given using natural units with $\hbar = c = 1$, where $\hbar$ is the reduced Planck constant and c is the speed of light in vacuum.

1.1 The Standard Model

The SM encompasses the theoretical foundations of particle physics. Since the model was formulated in the 1960s and 1970s [1–3], it has withstood a wide range of experimental tests [4] and is often described as one of the most successful scientific theories in history. The experimental confirmation of the model was completed in 2012 with the discovery of the Higgs boson by the A Toroidal LHC ApparatuS (ATLAS) and Compact Muon Solenoid (CMS) collaborations [5, 6].

The SM describes three out of the four fundamental forces of nature: the electromagnetic force, the strong force, and the weak force. The fourth fundamental force, gravity, is not included in the SM. However, gravity is much weaker than the other three forces and its effect is assumed too weak to notice on the small scales of particle physics.

1.1.1 Particle content

The elementary particles described by the SM can be grouped into the *fermions* and the *bosons*. The fermions have half-integer spin and obey Fermi-Dirac statistics, whereas the bosons have integer spin and obey Bose-Einstein statistics. An overview of the particle content of the SM is shown in figure 1.1, along with the mass and electric charge of each particle.

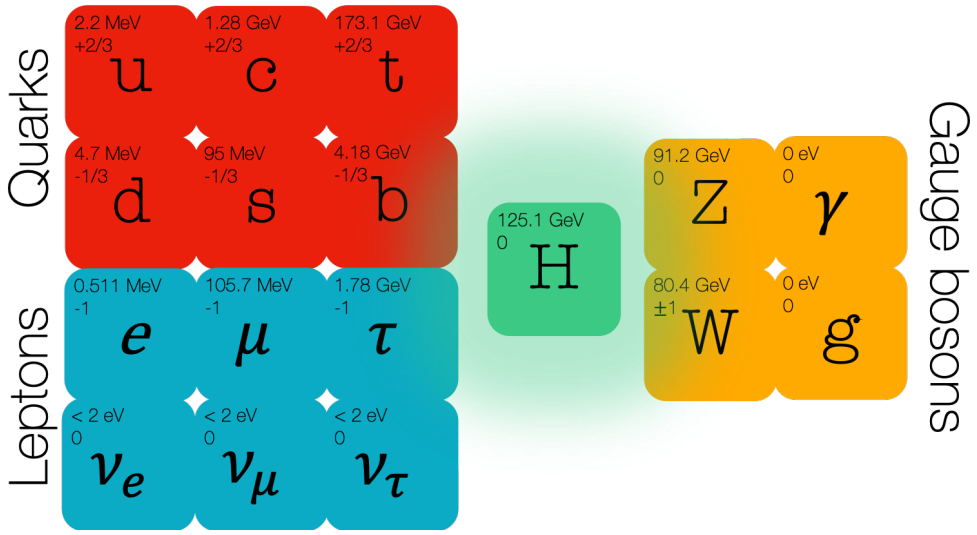


Figure 1.1: Overview of the fundamental particles in the Standard Model. The numbers in the top left corner of each box indicate the mass (top) and electric charge (bottom) of the particle (values taken from [4]). The quarks and leptons all have spin 1/2, the gauge bosons have spin 1, and the Higgs boson has spin 0.

The fermion group is composed of the quarks and the leptons, each of which are arranged in three generations which are distinguished by the masses of the particles. The quarks come in six flavours: the up (u), charm (c), and top (t) quarks carry electric charge +2/3 whereas the down (d), strange (s), and bottom (b) quarks carry electric charge -1/3. There are three charged leptons with different flavours, each carrying electric charge -1: the electron (e), muon (μ), and tau (τ). The three remaining leptons are the electrically neutral neutrinos, which come in the same flavours as the charged leptons: the electron neutrino (ν_e), muon neutrino (ν_μ), and tau neutrino (ν_τ).

The electron, up quark, and down quark compose the stable matter we observe in the world around us. Composite particles made up of two or more quarks are called *hadrons* and they are divided into the *mesons* and the *baryons*. The mesons are made up of an even number of quarks, such as the pions (π^0 is made up of $u\bar{u}$). Baryons consist of an odd number of quarks, the most common examples being the proton (uud) and neutron (udd).

The bosons are the force carriers that mediate the interactions between fermions. The electromagnetic force, felt only be electrically charged particles, is mediated by the massless and neutrally charged photon (γ). The strong force acts only on the quarks because these are the only fermions that carry colour charge. The colour charge comes in (anti-)red, (anti-)blue, and (anti-)green. Quarks carry one of these colours whereas gluons (g) carry a combination of one colour and one anti-colour. The strong interaction is mediated by eight massless gluons which have no electric charge. All particles mentioned above interact via the weak force which is mediated by the three massive gauge bosons: W^+ (electric charge +1), W^- (electric charge -1), and Z^0 (electric charge 0).

1.1.2 Symmetries

The SM is a renormalisable quantum field theory based upon gauge symmetries. Renormalisation refers to the procedure in which divergent parts of the perturbative calculations are absorbed by redefining certain observables, thereby leading to a finite result. The fundamental objects of the SM theory are the quantum fields which are defined at all points in space and time. The elementary particles are *excitations* of their corresponding fields. The symmetry group of the SM is

$$\mathrm{SU}(3)_C \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_{Y_w}, \tag{1.1}$$

where C stands for colour charge, L specifies that this symmetry only applies to fields with left-handed chirality, and Y_w is the weak hypercharge. The colour charge is associated to the SU(3) group and comes from the strong interaction which is described by Quantum Chromodynamics (QCD). The weak hypercharge is related to the electric charge, Q, and the third component of the weak isospin, T_3, as $Y_w = 2(Q - T_3)$. The electric charge is associated to the U(1) group and comes from the electromagnetic interaction which is described by Quantum Electrodynamics (QED). The weak isospin is associated to the group SU(2) and its third component, T_3, is the projection of the weak isospin along the z-axis. The product $\mathrm{SU}(2)_L \times \mathrm{U}(1)_{Y_w}$ describes the unification of the electromagnetic and weak forces in the *electroweak* theory. The quantum numbers Q, C, and T_3 are conserved in all SM interactions.

The full SM lagrangian consists of terms describing each of the three forces it encompasses, with the electromagnetic and weak forces combined in the electroweak theory. There are also terms relating to the Higgs mechanism. The full SM lagrangian is given by:

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{EW}} + \mathcal{L}_{\mathrm{QCD}} + \mathcal{L}_{\mathrm{Higgs}}. \tag{1.2}$$

Each term in this lagrangian respects the gauge invariances of the underlying symmetry group.

1.1.3 Electromagnetic interaction

The electromagnetic interaction has an infinite range and is described by QED. It originates from the $\mathrm{U}(1)_{Y_w}$ symmetry group which has one generator and therefore predicts one force mediator for this interaction. The U(1) symmetry means that QED is invariant under global gauge transformations of the form

$$\psi \to \psi' = e^{iQ\theta}\psi, \tag{1.3}$$

where ψ is a spin-1/2 field representing a fermion and Q is the electric charge. The spin-1/2 particles satisfy the Dirac equation of motion:

$$i\gamma^\mu \partial_\mu \psi - m\psi = 0, \tag{1.4}$$

where γ_μ are the Dirac gamma matrices and ∂_μ is the derivative with respect to x_μ. Once we require the theory to be invariant also under *local* gauge transformations ($\theta \to \theta(x)$), we need

to replace the standard derivative with the covariant derivative:

$$\partial_\mu \rightarrow D_\mu \equiv \partial_\mu - ieQA_\mu. \tag{1.5}$$

We see that a new field, A_μ, needs to be introduced in order to keep the lagrangian locally gauge invariant. The A_μ field is a spin-1 field representing the photon. The QED theory thus describes the interactions between the fermions and the photon. Only electrically charged fermions are affected by the electromagnetic force. The interaction vertex of the QED theory is shown in figure 1.2.

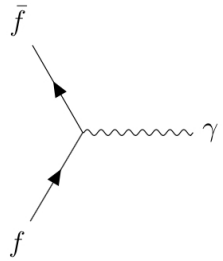


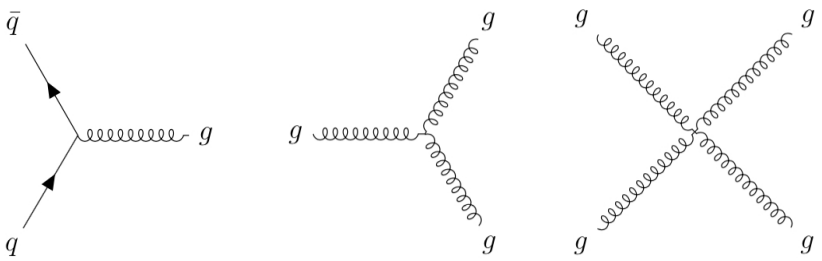
Figure 1.2: The QED interaction vertex.

1.1.4 Strong interaction

The strong interaction is described by QCD and originates from the SU(3)$_C$ symmetry which predicts eight force mediators for this interaction. Just like in QED, once we require invariance under local gauge transformations we need to introduce the covariant derivative:

$$\partial_\mu \rightarrow D_\mu \equiv \partial_\mu - ig_s T_a G^a_\mu, \tag{1.6}$$

where g_s is the strong coupling constant (usually referred to as $\alpha_s \equiv g_s^2/4\pi$), G^a_μ (with $a = 1-8$) are the gluon fields, and $T_a = \frac{1}{2}\lambda_a$ are the eight SU(3) generators with λ_a the Gell-Mann matrices. We see that we have now had to introduce eight new fields, G^a_μ, in order to retain local gauge invariance of the theory. The QCD theory describes the interactions between these eight gluons and the colour-charged quarks. Since the gluons carry colour charge as well, they can also interact amongst themselves. The quark-gluon interaction vertex and the gluon self-couplings are shown in figure 1.3.



(a) (b) (c)

Figure 1.3: The QCD interaction vertex (a), the triple gluon self-coupling (b), and the quartic gluon self-coupling (c).

The strong coupling constant α_s is not really a constant but is dependent on the separation between the particles involved. It is therefore also called the *running coupling*. The running coupling asymptotically diverges at large distance (or, equivalently, at low energies) which means that the strong interaction between two particles grows as they are pulled further apart. This leads to two interesting consequences named *asymptotic freedom* and *colour confinement*. The asymptotic freedom of QCD states that, inside hadrons, quarks and gluons roam about nearly freely. This is due to the fact that, at these very short distances, the strong interaction coupling is so weak that the partons hardly interact. This makes it possible to model quarks as free particles at high energies and therefore allows for the use of perturbative calculations to make very accurate predictions.

Colour confinement specifies that quarks and gluons do not exist in isolation but always form colourless hadrons. When a quark-antiquark pair is pulled apart, the energy of the strong interaction field between them increases due to the running coupling. Eventually, it becomes energetically favourable to create an additional quark-antiquark pair, rather than pulling apart the individual quarks any further. This then results in the formation of colourless hadrons before the quarks can be observed in isolation. The exception to this confinement is the top quark which is so heavy that it decays before it has had time to hadronise. Due to colour confinement, the range of the strong interaction is very small, of the order 10^{-15} m.

1.1.5 Electroweak interaction

The weak interaction originates from the $\mathrm{SU}(2)_L$ symmetry. The SU(2) group has three generators and thus predicts three mediating gauge bosons for this interaction, consisting of the $W^{\pm}$ and Z^0 bosons. The subscript L refers to the fact that the weak current only couples to particles with left-handed chirality and anti-particles with right-handed chirality. The weak interaction has the smallest range of all the forces of about 10^{-18} m.

The weak interaction consists of the charged and neutral currents, mediated by the $W^{\pm}$ and Z^0 bosons respectively. The charged current is the only interaction in the SM that can change the flavour of quarks. For example, in the beta decay of the neutron one of the down quarks in the neutron converts to an up quark to form the proton, while emitting a W boson that decays into a lepton-neutrino pair. The flavour changing of quarks does not only occur *within* generations but also *across* them. The probability of a transition from one flavour to another is encoded in the Cabibbo-Kobayashi-Maskawa (CKM) matrix [7, 8]. This matrix, V_{CKM}, relates the quark mass eigenstates to their flavour eigenstates by

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{\mathrm{CKM}} \begin{pmatrix} d \\ s \\ b \end{pmatrix} = \begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{1.7}$$

where the *d*, *s*, and *b* are the quark mass eigenstates and *d'*, *s'*, and *b'* are the flavour eigenstates. The diagonal elements of this matrix are all very close to one which means that the transitioning

of quark flavours happens most often within a generation. However, the off-diagonal elements cannot be neglected, with $|V_{us}|$ and $|V_{cd}|$ being approximately 0.2 [4].

The weak force was shown to be unifiable with the electromagnetic force by Glashow, Weinberg, and Salam [1–3]. This means that these two forces can be considered as two manifestations of the same fundamental interaction. Below the unification energy ($\approx$ 246 GeV), the two forces can be identified as two separate interactions, while above this unification scale they merge into one. The symmetry group underlying the electroweak theory is $\mathrm{SU}(2)_L \times \mathrm{U}(1)_{Y_w}$. The $\mathrm{SU}(2)_L$ group has three generators and the $\mathrm{U}(1)_{Y_w}$ group has one. This leads to four gauge bosons involved in the electroweak interaction. The Feynman diagrams describing the interaction vertices and the self-couplings of the weak gauge bosons are shown in figure 1.4.

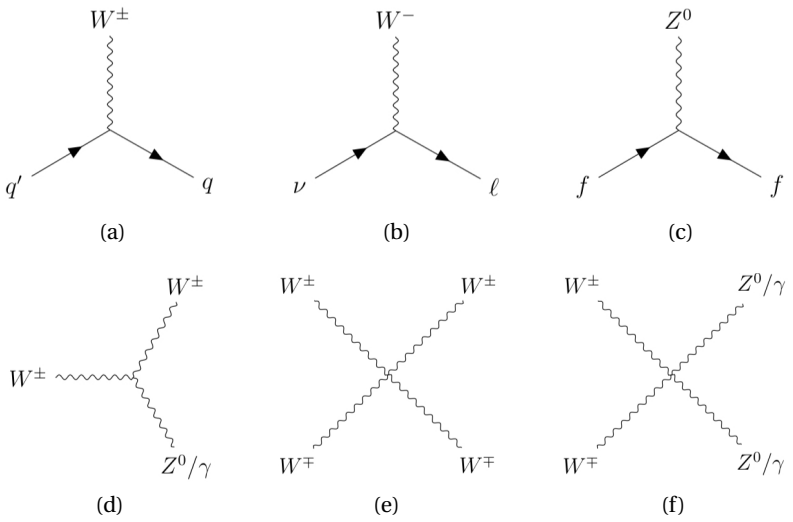


Figure 1.4: Examples of the electroweak interaction vertices for the charged currents in (a) and (b) and for the neutral current in (c). The electroweak self-couplings are shown in (d), (e), and (f).

The SM electroweak lagrangian is obtained by requiring local gauge invariance by introducing the covariant derivative:

$$\partial_\mu \rightarrow D_\mu \equiv \partial_\mu - i g \vec{T} \cdot \vec{W}_\mu - i g' \frac{Y}{2} B_\mu, \tag{1.8}$$

where g and g' are the coupling constants of the $\mathrm{SU}(2)_L$ and $\mathrm{U}(1)_{Y_w}$ groups, respectively. The $\vec{W}_\mu$ and B_μ are the gauge fields of the symmetry groups. These gauge boson fields are required to be massless since adding a mass term to the Lagrangian 'by hand' violates gauge invariance. However, we know from experiment that the $W^\pm$ and Z^0 bosons are massive. Therefore, another method was introduced by Englert, Brout, and Higgs in 1964 [9, 10] by which these bosons acquire their mass, known as the *Higgs mechanism*.

1.2 The Higgs mechanism

In order to allow for massive vector bosons in the electroweak theory while keeping gauge invariance intact, a complex scalar field, ϕ, is introduced. This field has a scalar potential of the form

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2, \tag{1.9}$$

which is invariant under the SU(2) × U(1) symmetry group. The first term in the potential indicates the scalar mass term whereas the second term represents the self-interaction vertex. We require that $\lambda > 0$ such that the potential energy is bounded from below, but the parameter μ can be chosen freely. If we choose $\mu^2 > 0$, both terms in the potential are positive and $V(\phi)$ takes on a parabolic shape with a unique minimum at $\phi_0 = 0$. However, if we choose $\mu^2 < 0$, the potential takes on the shape of a 'Mexican hat', as shown in figure 1.5. Here, we see that $\phi = 0$ represents a local minimum, but the global minimum is not unique: any point around the circular base of the 'hat' can be chosen as the minimum. This circle of minima is given by

$$\phi_0 = \sqrt{-\frac{\mu^2}{2\lambda}} = \sqrt{\frac{v^2}{2}}, \tag{1.10}$$

where $v = \sqrt{-\mu^2/\lambda}$ is the *vacuum expectation value* of the Higgs field, which is now non-zero. The value of v is given by $v = \sqrt{1/\sqrt{2}G_F} \approx 246$ GeV, where G_F is the Fermi constant.

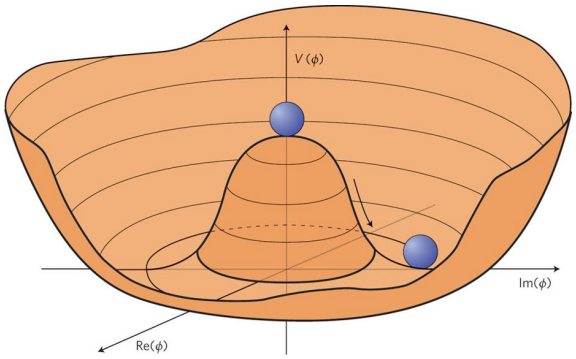


Figure 1.5: Illustration of the Higgs potential which takes on the shape of a 'Mexican hat' [11]. The vacuum can be picked from any point around the bottom of the hat which leads to spontaneous symmetry breaking.

The choice of one of the infinite possible minima as the ground state of ϕ will result in the spontaneous breaking of the $SU(2)_L \times U(1)_{Y_w}$ symmetry to $U(1)_{QED}$. Evidently, the vacuum is not invariant under the full electroweak symmetry group. It is through this spontaneous symmetry breaking that the gauge bosons and fermions acquire their mass through the interaction with the Higgs field.

The introduction of a complex scalar field in the SM adds four additional degrees of freedom to the theory. Three of these degrees of freedom are manifested as *Goldstone bosons*. The Goldstone theorem [12] states that, for every broken continuous symmetry, a massless scalar particle appears. These Goldstone bosons can be absorbed by a gauge field which leads to massive gauge bosons with an extra longitudinal polarisation component. In the case of elec-

troweak symmetry breaking, the three generators of the broken SU(2) group lead to three Goldstone bosons. These are absorbed by the weak gauge fields and thereby generate the masses of the $W^{\pm}$ and Z^0 bosons. These masses can be obtained from the Higgs lagrangian in which the W^1_μ and W^2_μ fields from equation 1.8 are combined to make the $W^{\pm}$ bosons and the W^3_μ and B_μ fields mix to make the Z^0 boson and the photon. The tree level predictions for their masses are

$$m_W = \frac{vg}{2}, \quad m_Z = v\frac{\sqrt{g^2+g'^2}}{2}, \quad m_\gamma = 0, \tag{1.11}$$

where g and g' are the coupling constants of the $SU(2)_L$ and $U(1)_{Y_w}$ groups, respectively.

The fourth degree of freedom from the Higgs complex scalar field forms a quanta of the field called the *Higgs boson*. This boson has a mass of

$$m_H = \sqrt{-2\mu^2} = \sqrt{2\lambda v^2}. \tag{1.12}$$

Since λ is not predicted by the theory, the mass of the Higgs boson can only be found experimentally.

When we write out the full Higgs lagrangian (all terms including ϕ which are invariant under $SU(2)_L \times U(1)_{Y_w}$), we find that the Higgs field couples to fermion fields with coupling strengths y_f called the *Yukawa couplings*. Due to this coupling, the fermions can acquire mass whilst also preserving the gauge invariance of the weak interaction. The tree level prediction for the fermion masses is related to the Yukawa coupling and vacuum expectation of the Higgs field through

$$m_f = y_f \frac{v}{\sqrt{2}}, \tag{1.13}$$

where f stands for any fermion in the SM. The Yukawa couplings are not predicted, thus the fermion masses need to be determined experimentally. Measurements of the couplings between the Higgs and fermions are therefore important in order to validate the fermion mass relation in equation 1.13. The interactions between the Higgs boson and weak gauge bosons as well as fermions are shown in figure 1.6. The Higgs boson also interacts with itself via a triple and quartic coupling.

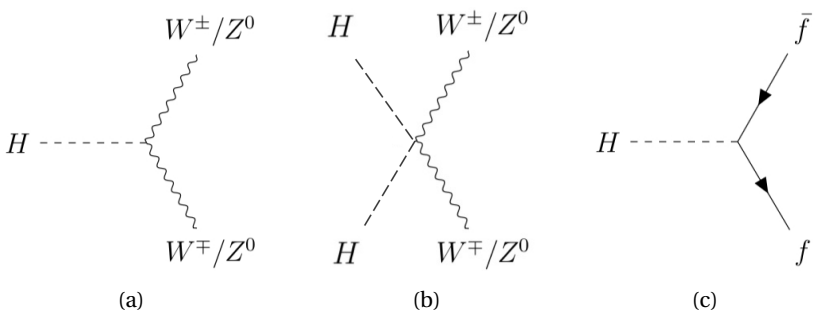


Figure 1.6: The interaction vertices of the Higgs coupling to the weak gauge bosons, (a) and (b), and to fermions (c).

1.3 The Higgs boson

The SM Higgs boson is a particle with a spin and electric charge of 0. Its mass, given by equation 1.12, is a free parameter of the SM theory and thus needs to be established in experiment. The mass has been measured in various analyses and is found to be 125.18 ± 0.16 GeV [4]. The time between the Higgs boson's first theoretical description by Brout, Englert, and Higgs [9, 10] in 1964 and its experimental discovery in 2012 by the ATLAS [5] and CMS [6] collaborations was the longest of all the fundamental particles, as shown in figure 1.7. The Higgs boson discovery was a very important milestone for particle physics, since it completed the experimental observation of the particle content of the SM. The SM predicts many ways in which the Higgs boson can be produced and subsequently decay. All of these production and decay modes need to be studied in order to determine the properties of this recently discovered particle. The various modes are discussed in this section.

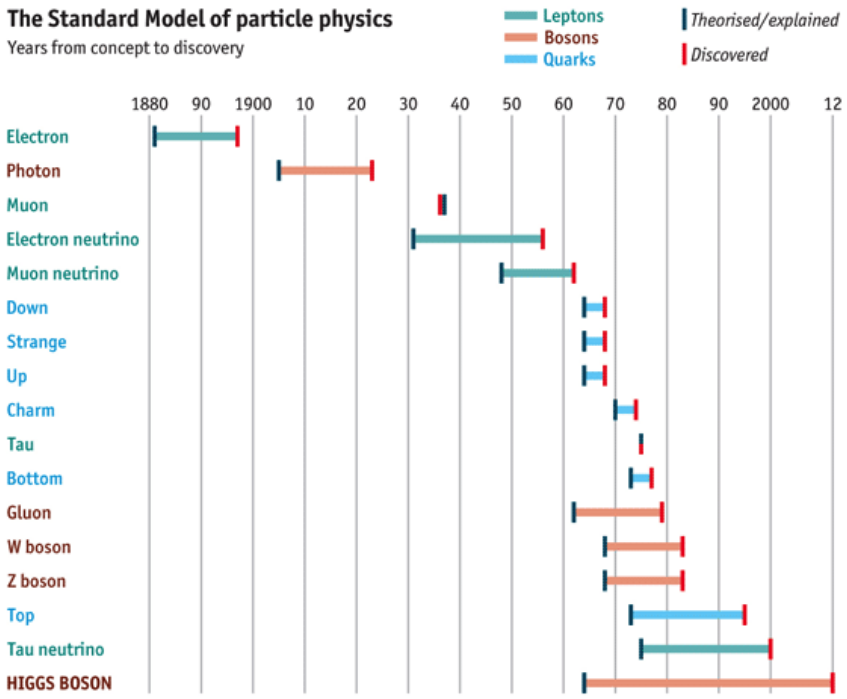


Figure 1.7: Overview of the number of years between theoretical concept of the fundamental particles and their experimental observation [13].

1.3.1 Higgs boson production modes

An overview of the cross-sections of the main production modes of the Higgs boson at the Large Hadron Collider (LHC) as a function of the centre-of-mass energy is shown in figure 1.8. At the current energy of $\sqrt{s} = 13$ TeV, the four main production modes are gluon gluon fusion (ggF), vector boson fusion (VBF), associated vector boson production (VH), and associated top quark pair production ($t\bar{t}H$). Exemplary Feynman diagrams of these production modes are shown in figure 1.9.

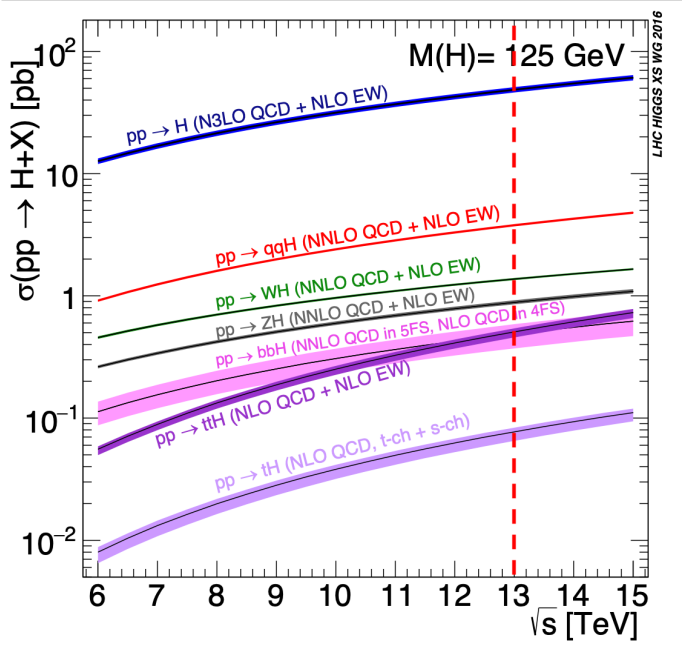
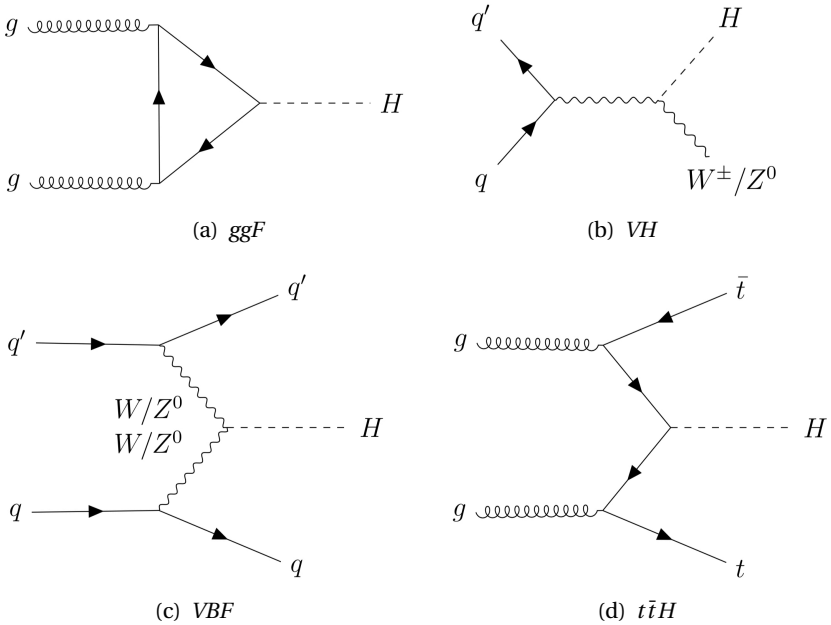


Figure 1.8: The cross-sections of the main production modes of the Standard Model Higgs boson and their uncertainties as a function of the centre-of-mass energy of the LHC [14]. The current energy of 13 TeV is indicated with the red dashed line.



(a) *ggF* (b) *VH*

(c) *VBF* (d) $t\bar{t}H$

Figure 1.9: Exemplary Feynman diagrams of the four main production modes of the Higgs boson at the LHC.

The *VH* and *VBF* processes probe the coupling of the Higgs boson to the weak gauge bosons, whereas the *ggF* and $t\bar{t}H$ processes both probe the Yukawa couplings between the Higgs and the quarks. The *ggF* production mode occurs about 100 times more frequently than the $t\bar{t}H$ mode, with cross-sections of 48.6 pb and 0.50 pb respectively [4]. This makes the *ggF* process a good candidate to study the Higgs couplings to fermions. In the *ggF* process, the gluons couple to the Higgs via a virtual quark loop which mainly involves top quarks because the Yukawa coupling strength is proportional to the quark mass (see equation 1.13). It can thus be used to study the coupling of the heaviest fermion, the top quark, to the Higgs boson. However, this measurement makes the assumption that there are no Beyond the Standard Model (BSM) effects in the loop. The $t\bar{t}H$ process gives tree-level access to the top Yukawa coupling and is thus a good alternative way to measure this coupling while significantly reducing the model dependence of the measurement.

1.3.2 Higgs boson decay modes

The Higgs boson has a lifetime of about 10^{-22} s and is therefore not observed directly. We study its decay products which can consist of a wide variety of particles, as shown in figure 1.10. For the measured Higgs mass of 125 GeV, the boson decays most frequently to a pair of bottom quarks. The exact branching ratios of the various decay modes are listed in table 1.1. Since the Higgs boson does not couple to massless particles, the $H \to \gamma\gamma$ and $H \to gg$ decay modes are induced through loops of massive particles.

The Higgs discovery in 2012 was most reliant on the ZZ and $\gamma\gamma$ decay channels [5, 6]. The first observation of fermionic decay of the Higgs boson was to a pair of tau leptons in 2016 [15], followed by observation of the $H \to b\bar{b}$ decay mode in 2018 [16, 17].

Decay channel	**Branching ratio [%]**
$H \to b\bar{b}$	58.2
$H \to WW$	21.4
$H \to gg$	8.19
$H \to \tau\tau$	6.27
$H \to c\bar{c}$	2.89
$H \to ZZ$	2.62
$H \to \gamma\gamma$	0.227
$H \to Z\gamma$	0.153
$H \to \mu\mu$	0.022

Table 1.1: The expected branching ratios of the Standard Model Higgs boson with a mass of 125 GeV [14].

1.4 The top quark

The top quark was the last of the quarks to be observed, with a joint discovery by the CDF and D0 collaborations in 1995 [18, 19]. With a mass of 173.1±0.9 GeV [4], the top quark is the

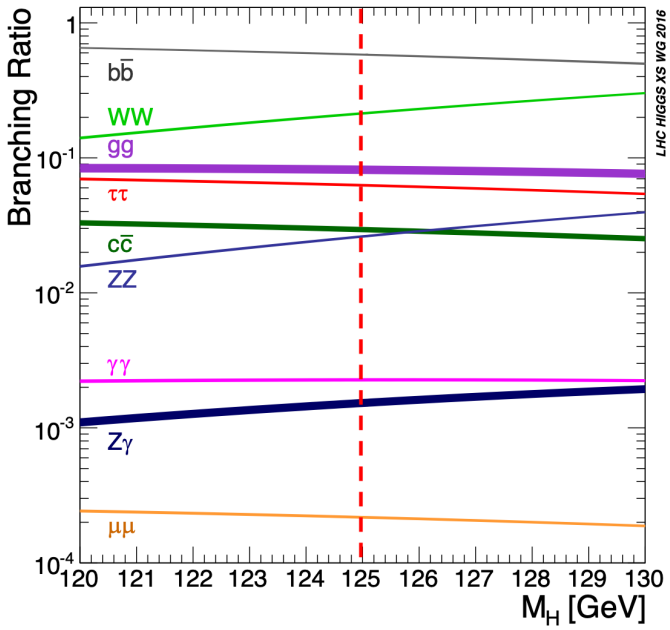


Figure 1.10: The predicted branching ratios of the main decays of the Standard Model Higgs boson and their uncertainties as a function of the Higgs boson mass [14]. The experimental value of $M_{\mathrm{H}} \approx 125$ GeV is indicated with the red dashed line.

heaviest particle in the SM. It has an electric charge of +2/3 and, since it is a fermion, a spin of 1/2. The high mass of the top quark indicates a strong Yukawa coupling to the Higgs boson:

$$y_t = \frac{\sqrt{2}m_t}{v} \approx 1, \tag{1.14}$$

where v is the vacuum expectation value of the Higgs field.

1.4.1 Top quark production modes

At the LHC, the most common way to produce top quarks is through top quark pair production. This occurs via quark-antiquark annihilation or gluon-gluon fusion, where the latter is dominant due to the large amount of gluons produced in the proton-proton collisions. For a top quark mass of 172.5 GeV, the cross-section of $t\bar{t}$ at a centre-of-mass energy of 13 TeV is 832^{+46}_{-51} pb [20–23].

1.4.2 Top quark decay modes

The top is the heaviest particle in the SM with a lifetime of about 10^{-25} s [4] which means it decays before it has a chance to hadronise. This makes the top quark unique and means it can decay into a W boson and a lighter quark. From equation 1.7, we note that the top quark can transfer into a down, strange, or bottom quark via the weak interaction. However, the CKM matrix elements $|V_{td}|$ (≈ 0.009) and $|V_{ts}|$ (≈ 0.04) are negligible compared to $|V_{tb}|$ which is almost equal to one [4]. Therefore, the only significant decay mode is $t \to W^+ + b$.

The decay of top quarks is categorised according to the decay products of the W boson. A hadronic top has the W boson decaying to a quark pair whereas a leptonic top has the W decaying to a lepton and its corresponding neutrino. Events of top quark pairs, $t\bar{t}$, are classified as fully hadronic, semileptonic, or dileptonic depending on whether none, one, or both of the tops decay leptonically. The branching ratios of the various $t\bar{t}$ decay modes are displayed in figure 1.11. This chart shows that the hadronic, semileptonic, and dileptonic decay modes contribute 46%, 30%, and 4% respectively. This excludes the decays involving τ leptons which account for 20% of the $t\bar{t}$ decays. These decay modes are usually considered separately, depending on the consecutive τ decay. If the τ decays leptonically, the top is also classified as leptonic and if the τ decays hadronically, the tau is reconstructed and the top is classified as hadronic.

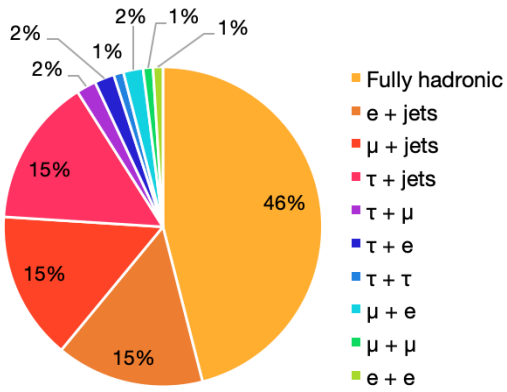
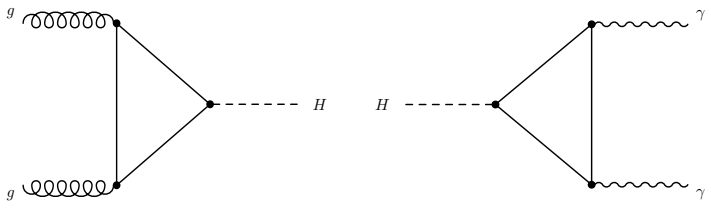


Figure 1.11: Pie chart illustrating the branching ratios in % of each of the $t\bar{t}$ decay modes.

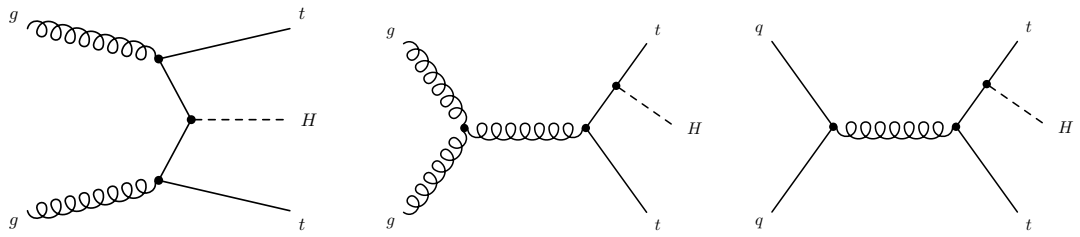
1.5 Top-Higgs Yukawa coupling

As mentioned above, the top quark Yukawa coupling can be probed with the *ggF* Higgs boson production mode (see figure 1.12 (a)). The $H \to \gamma\gamma$ decay mode provides another probe into this coupling since it occurs via a heavy-particle loop dominated by top quarks (see figure 1.12 (b)). However, both of these processes occur via loops and the measurements assume no BSM effects. The $t\bar{t}H$ process is an alternative way to measure the top-Higgs Yukawa coupling while significantly reducing the model dependence of the measurement. This process can be induced through two quarks or two gluons, as shown in figure 1.13.



(a) (b)

Figure 1.12: Coupling of the Higgs boson to the top quark in the *ggF* production mode (a) and $H \to \gamma\gamma$ decay mode (b).



(a) (b) (c)

Figure 1.13: Exemplary leading order Feynman diagrams of $t\bar{t}H$ production.

The $t\bar{t}H$ process allows for a direct tree-level measurement of the top Yukawa coupling which serves as a high precision test of the SM. Any deviations in y_t from equation 1.13 would indicate new physics. The tH process also provides direct access to this coupling but is highly suppressed compared to $t\bar{t}H$ at a centre-of-mass energy of 13 TeV (see figure 1.8). The loop and tree-level measurements are complimentary and can be compared in order to probe the presence of BSM effects in the top quark loop. The current best measurement of the top Yukawa coupling comes from a combination of Higgs boson measurements from both ATLAS and CMS using the Run I dataset collected at $\sqrt{s} = 7$ and 8 TeV. The measured value of y_t is 0.87 ± 0.15 times the SM prediction, assuming that no BSM particles couple to the Higgs boson in the *ggF* and $H \to \gamma\gamma$ loops [15].

1.5.1 $t\bar{t}H$

This thesis discusses the search for the $t\bar{t}H$ production mode in ATLAS with 13 TeV data of Run II of the LHC. Focus is given to the $H \to b\bar{b}$ decay channel. The current search builds on previous searches for the same process performed with ATLAS data recorded at $\sqrt{s} = 7$ TeV [24] and 8 TeV [25, 26]. The results of these searches are expressed in terms of the signal strength parameter

$$\mu_{t\bar{t}H} = \frac{\sigma_{\text{observed}}}{\sigma_{\text{expected}}}, \tag{1.15}$$

where σ_{expected} is the SM cross-section. The signal strength found for the full combination of Run I $t\bar{t}H(H \to b\bar{b})$ searches in ATLAS is $\mu_{t\bar{t}H} = 1.4 \pm 1.0$ [26]. The corresponding sensitivity found by the CMS collaboration is 0.7 ± 1.9 [27]. The results from both experiments were combined for the $H \to b\bar{b}$ decay channel as well as other $t\bar{t}H$ decay channels to obtain a final $t\bar{t}H$ sensitivity in Run I of $\mu_{t\bar{t}H} = 2.3^{+0.7}_{-0.6}$ with an observed (expected) significance of $4.4\sigma(2.0\sigma)$ [15].

1.5.2 Boosted $t\bar{t}H$

At a centre-of-mass energy of 13 TeV, Higgs bosons can be produced with transverse momenta well above their rest mass, which means that they can be probed in the boosted (high-p_T) regime. This thesis studies the boosted $t\bar{t}H(H \to b\bar{b})$ process in the semileptonic decay channel of the top quark pair. The effect of applying a boost to the p_T of the Higgs boson and hadronically decaying top quark is shown in figure 1.14. When the boost is applied, the decay products

of these particles become highly collimated. A rule of thumb is that the decay products will be produced in a cone of $R \approx 2m/p_T$, where m is the mass and p_T the transverse momentum of the initial particle. In this way, many of the final-state jets can be combined together into large jets. This results in a simplified combinatorial background which makes it easier to reconstruct the objects in the events. In turn, this can help to increase the purity of signal regions and thereby improve the sensitivity of the analysis.

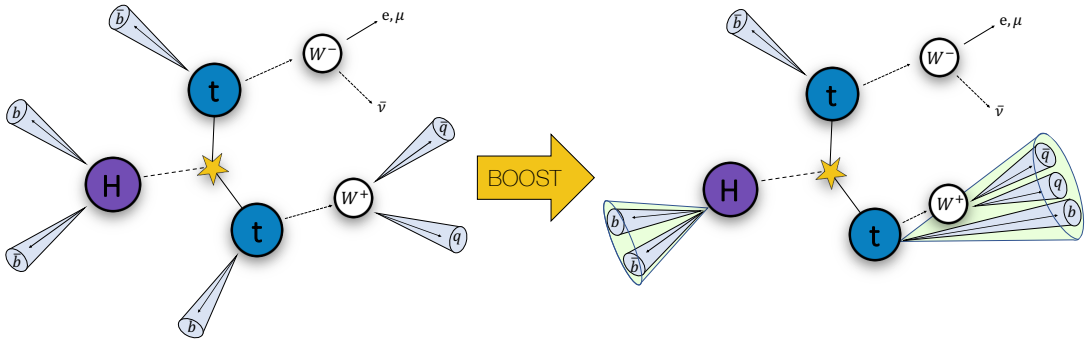


Figure 1.14: Illustration of the effect of applying a boost in p_T to the hadronic top and Higgs boson in a $t\bar{t}H(H \to b\bar{b})$ event.

2 THE ATLAS EXPERIMENT AT THE LHC

Founded in 1954, the European Organization for Nuclear Research (CERN) is the largest international particle physics laboratory in the world. It is located on the Franco-Swiss border near Geneva and houses the world's largest and most powerful particle accelerator: the LHC. The LHC is mainly designed to collide protons at a maximum centre-of-mass energy ($\sqrt{s}$) of 14 TeV. It also collides heavy ion beams for a short period each running year at lower energies.

The first concept for the LHC was drawn up in 1984 and in 2008 it had its first proton beam circling around the ring. Later in 2008, however, an electrical failure caused liquid helium (used for cooling the machine's magnets) to leak into the LHC tunnel. After a period of necessary repairs and upgrades to prevent a similar incident in the future, the LHC was once again circulating proton beams just over a year later. In 2009, *proton-proton (pp)* collisions were recorded at world record energies of 1.18 TeV and 2.36 TeV which marked the start of the LHC's first physics run (Run I). In 2010, *pp* collisions at $\sqrt{s}$ = 7 TeV were started which was increased to 8 TeV in 2012 until the end of Run I in early 2013. There was an extended shutdown between 2013 and 2015 to allow for machine upgrades to the LHC and the experiments it houses. The second run (Run II) lasted from 2015 to 2018 and provided $\sqrt{s}$ = 13 TeV *pp* collisions. The ATLAS detector is one of four experiments placed on the LHC ring and recording the data arising from the collisions. This chapter will give an overview of the LHC accelerator complex and the ATLAS detector.

2.1 The Large Hadron Collider at CERN

The accelerator consists of several stages, as shown in figure 2.1. The sequence of machines accelerates beams of protons to nearly the speed of light before they are injected into the final LHC ring which has a circumference of 27 km. The protons are sourced from a bottle of hydrogen gas, where the electrons are stripped off by an electric field. They are injected into the Linear Accelerator 2 (Linac 2) and linearly accelerated to an energy of 50 MeV. The energy is then increased to 1.4 GeV by the Booster Proton Synchrotron (booster in figure 2.1) and to 25 GeV by the Proton Synchrotron (PS in figure 2.1). The Proton Synchrotron divides the proton beams into small packets called *bunches*. Each bunch contains approximately 10^{11} protons and

they are spaced 25 ns apart in Run II. The bunches are accelerated further by the Super Proton Synchrotron (SPS) to 450 GeV. After the LHC is filled with the proton bunches it takes another 20 minutes for them to reach their final energy of 6.5 TeV [28]. Inside the LHC, the bunches are accelerated by an electric field at one location along the ring.

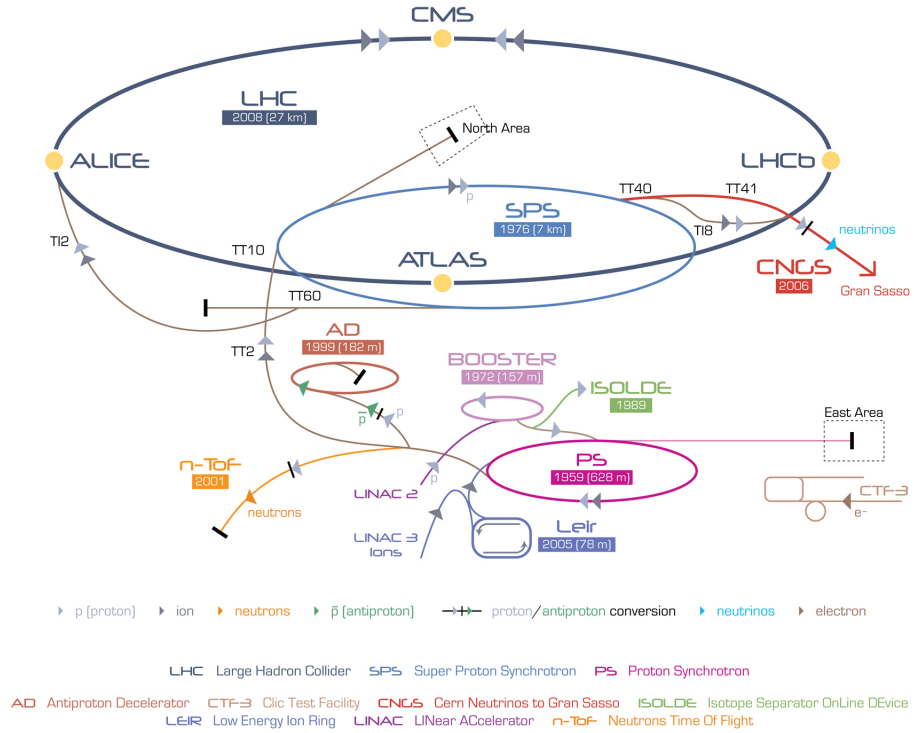


Figure 2.1: The LHC accelerator complex [29].

There are two proton beams going around the LHC ring in separate vacuum pipes and opposite directions. The curvature in the path of the proton beams is achieved by dipole magnets which exert a force perpendicular to the proton velocity through the Lorentz force. The superconducting coils in the dipole magnets create opposite polarity magnetic fields for each of the vacuum pipes which allows the two proton beams to travel in opposite directions. The beam is focused in width and height by the use of quadrupole magnets.

When the beams have reached their peak energy, they are brought together to collide at four interaction points. Dipole magnets deflect the beams towards the collision point, and eight sets of so-called *inner triplet* magnets focus the beams prior to the collision. These inner triplets consist of three superconducting quadrupole magnets. The four collision points house four different LHC experiments: ATLAS, A Large Ion Collider Experiment (ALICE), CMS, and Large Hadron Collider beauty (LHCb). The ALICE experiment is a heavy-ion detector and is designed to study strongly interacting matter in the quark-gluon plasma. CMS is a general-purpose detector built for similar physics goals as ATLAS which include SM measurements and BSM searches such as dark matter and supersymmetry (SUSY). The LHCb detector is specialised in b-physics and designed to measure charge-parity (CP) violation in order to understand the matter-antimatter asymmetry in the universe. There are three other experiments using the LHC accelerator which are smaller and more specialised and are not relevant for this work.

2.1.1 Luminosity

The experiments record the data from the proton-proton and heavy ion collisions. The number of collisions per area and per second is measured by the instantaneous luminosity:

$$\mathcal{L} = \frac{N_p^2 k_b f}{4\pi\sigma_x\sigma_y} F, \tag{2.1}$$

where N_p is the number of protons per bunch ($\sim 10^{11}$), k_b the number of bunches per beam (~ 2800), f the bunches' crossing frequency (maximum 40 MHz), σ the transverse size of the bunch at the interaction point ($\sim 16\mu$m), and F is a correction factor to account for the crossing angle of the proton beams at the interaction point (~ 0.7). The luminosity is related to the number of events, N, of a certain process, i, by

$$N_i = \sigma_i \int \mathcal{L}\, dt, \tag{2.2}$$

where σ_i is the cross-section of the process under investigation. The total amount of data recorded in the LHC experiments is given by the integral over the instantaneous luminosity: $L = \int \mathcal{L}\, dt$. To date, the LHC has delivered 156 fb^{-1} of data to the ATLAS experiment during Run II of which it has recorded 147 fb^{-1} (see figure 2.2). This corresponds to a recording efficiency of 94% which is due to the inefficiency of the data acquisition (DAQ) in ATLAS.

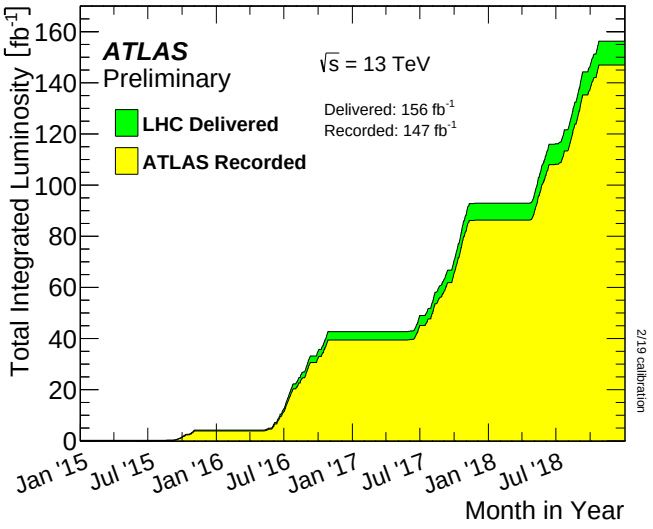


Figure 2.2: Total integrated luminosity at the end of Run II (December 2018) showing 13 TeV proton-proton data only [30].

2.1.2 Pile-up and underlying event

Each bunch that is filled into the LHC contains approximately 10^{11} protons and they are spaced 25 ns apart in Run II. These conditions lead to a very high number of *pp* interactions at every bunch crossing. The high-energy collision of interest is called the *hard scatter*. In Run II, we have dozens of additional interactions which are referred to as *pile-up*. We distinguish between in-time and out-of-time pile-up. The former refers to additional *pp* collisions in the same

bunch-crossing as the hard scatter we are interested in, whereas the latter indicates additional *pp* collisions occurring in neighbouring bunch-crossings, just before or after the hard scatter. Pile-up typically causes additional low-energy deposits in the ATLAS detector. Such additional deposits also originate from the underlying event (UE): soft additional jets produced in the same *pp* collision involved in the hard scatter.

The distribution of pile-up is determined by the mean number of interactions per bunch crossing as shown in figure 2.3. This mean corresponds to the mean of the Poisson distribution of the number of interactions per bunch crossing calculated for each bunch. It is calculated as $\mu = L_{\text{bunch}} \times \sigma_{\text{inel.}} / f_r$ where L_{bunch} is the instantaneous luminosity per bunch, σ_{inel} is the inelastic cross-section taken as 80 mb for 13 TeV collisions, and f_r is the LHC revolution frequency.

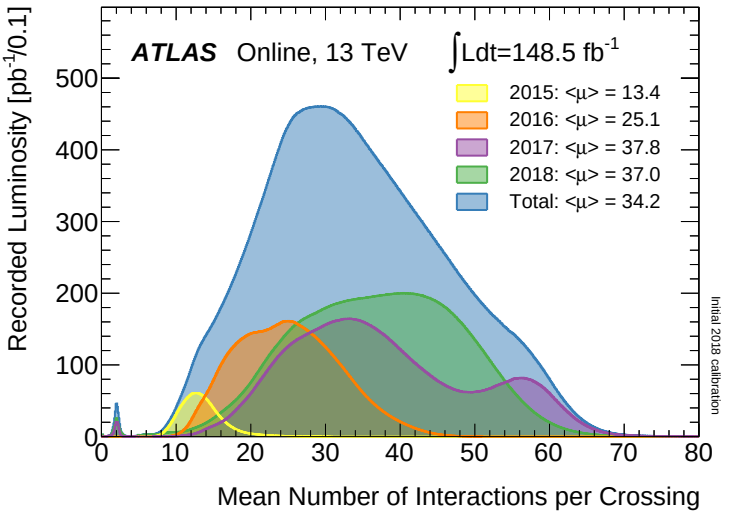


Figure 2.3: Distribution of the mean number of interactions per bunch crossing for proton-proton collision data at 13 TeV from 2015-2018 [30]. All data recorded by ATLAS during stable beams is shown, and the integrated luminosity and the mean μ value are given in the figure.

2.2 The ATLAS detector

The ATLAS detector [31] is located at interaction point 1 of the LHC. It is a general purpose experiment built to perform Standard Model measurements and searches for physics beyond the Standard Model. The experiment consists of several complementary subdetectors as shown in figure 2.4. The inner part of the detector is surrounded by a 2 T solenoidal magnetic field and a 4 T toroidal field is present outside of the calorimeters, surrounding the muon system. The subdetectors and magnet system are described in more detail in the following sections.

The ATLAS coordinate system (shown in figure 2.5) is a Cartesian right-handed system with the nominal interaction point defined as the origin. The z-axis lies along the beam direction and the $x-y$ plane is transverse to the beam. The transverse-momentum (p_T), -energy $\left(E_T = \sqrt{m^2 + p_T^2}\right)$, and -missing energy ($E_T^{\text{miss}}$) are defined in the $x-y$ plane. The azimuthal

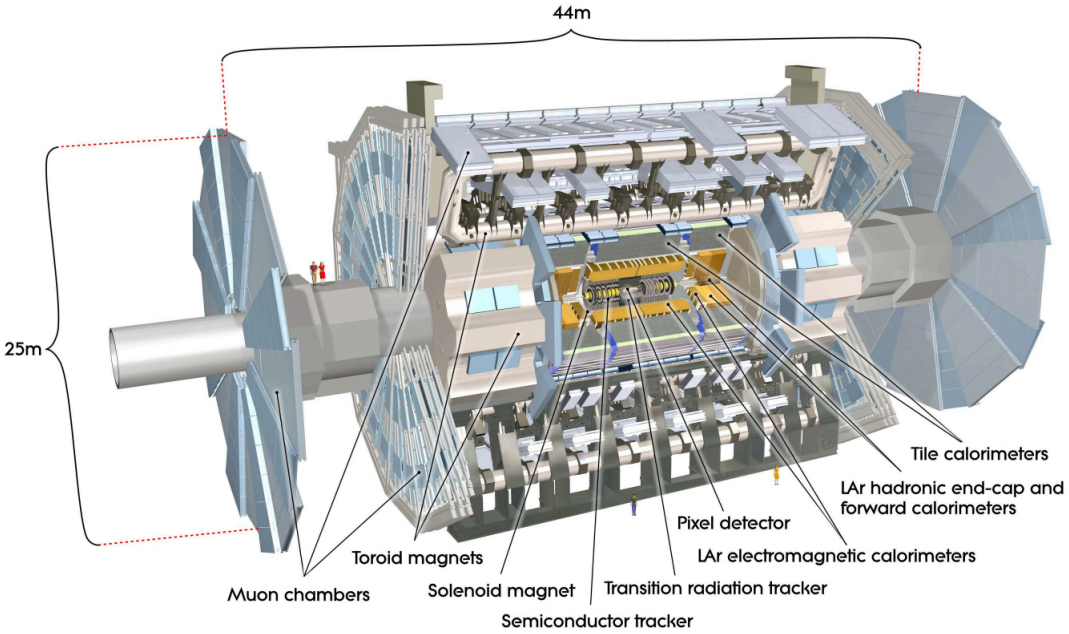


Figure 2.4: The ATLAS detector and its sub-systems [31].

angle ($0 \leq \phi \leq 2\pi$) is measured around the beam axis and the polar angle ($0 \leq \theta \leq \pi$) is measured from the beam axis. Instead of using the polar angle directly, it is convenient to define the rapidity $y = \frac{1}{2} \ln \left[\frac{E+p_z}{E-p_z} \right]$ because differences in this parameter are invariant under boosts along the beam axis. The rapidity is usually approximated by the pseudorapidity η in the high energy limit ($E \gg mc^2$) in order to have a purely angular definition which is easier to interpret in the detector. The rapidity and pseudorapidity are equal for massless particles. The pseudorapidity is related to the polar angle by:

$$\eta = -\ln \left[\tan \frac{\theta}{2} \right]. \tag{2.3}$$

The pseudorapidity is zero along the y-axis ($\theta = 90°$) and 4.74 close to the beam axis ($\theta = 1°$). The central detector region ($|\eta| < 1.5$) is referred to as the *barrel*, the region contained in $1.5 < |\eta| < 2.5$ as the *end-cap*, and any larger pseudorapidity refers to the *forward* regions of the detector. We define the distance parameter ΔR measuring the distance between objects in η, ϕ space as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

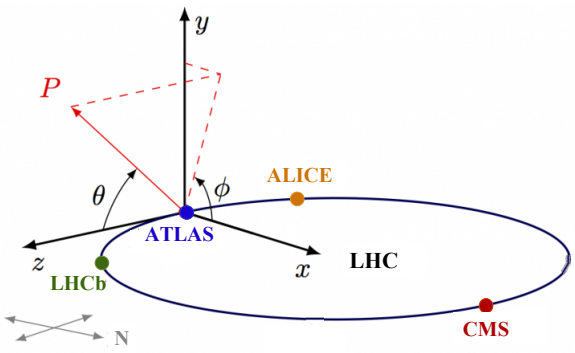


Figure 2.5: The ATLAS coordinate system, picture adapted from [32].

2.2.1 The magnet system

We make use of the Lorentz force on charged particles in order to measure their momenta. To this end, a magnetic field that bends the charged particles' trajectories is applied in two parts of the detector. The direction and amount of track deflection tells us the momentum of the particle. The ATLAS magnet system is composed of four superconducting magnets which provide a magnetic field mostly orthogonal to the particle trajectories. The tracking detector, also called the inner detector (ID) because it is closest to the beam pipe, is surrounded by a solenoid which provides a field parallel to the beam axis. The other three magnets are superconducting toroidal magnets embedded into the outermost detector layer: the muon spectrometer. In order to keep the magnets superconducting, they need to be cooled down to $\sim$ 4.5 K which is accomplished by using liquid helium. A schematic view of the magnet system is shown in figure 2.6.

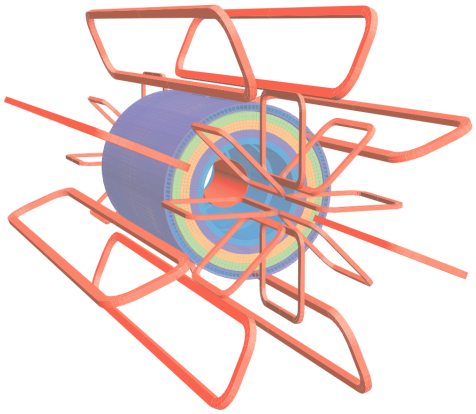


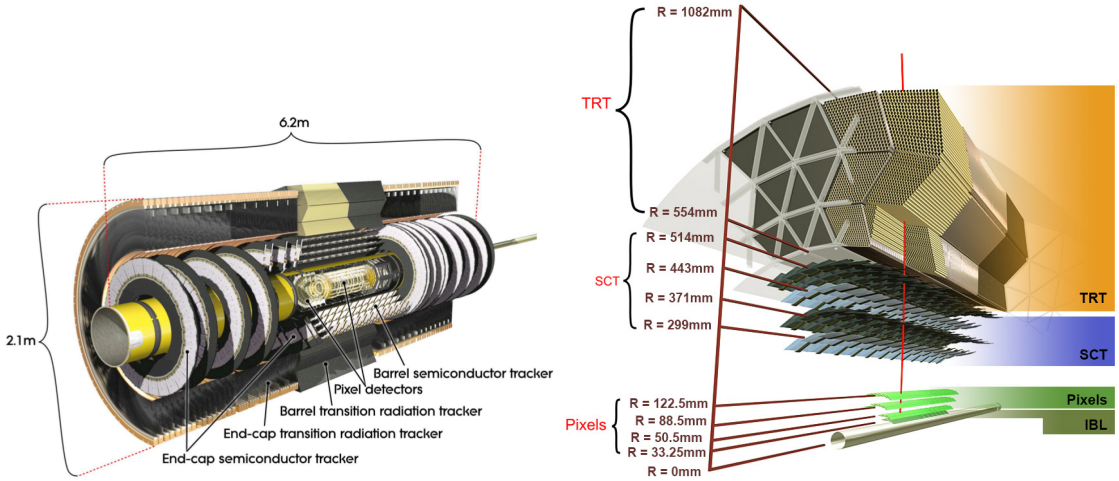
Figure 2.6: The ATLAS magnet system showing the three toroids in red and the central solenoid enclosed by the calorimeter layers [31].

The solenoid magnet is designed to produce an axial magnetic field of 2 T and is located between the inner detector and calorimeters. In order to have the best possible calorimeter performance with a minimal level of particle interactions in the solenoid coil, some design constraints had to be met in order to keep the material thickness as low as possible. The single-layer coil is wound with an aluminium-stabilised niobium-titanium conductor.

There are two end-cap toroid magnets and one barrel toroid which are each composed of eight coils as shown in figure 2.6. The toroid coils use an aluminium-stabilised niobium-titanium-copper conductor. The end-cap magnets produce a peak field of 4.1 T whereas the barrel produces 3.9 T. The toroid magnets do not cover the full solid angle because the use of less material reduces multiple scattering effects which are hard to reconstruct and degrade the muon momentum resolution.

2.2.2 The inner detector

The ATLAS ID is used for the reconstruction of the paths of charged particles (*tracking*), reconstruction of interaction vertices, and the identification of electrons and positrons. The ID is immersed in a 2 T solenoidal magnetic field which allows for the extraction of the momentum of charged particles. A very high precision is needed in order to distinguish between the ~ 1000 particles that emerge from the collision point every 25 ns in Run II. The fine detector granularity needed for this is achieved by silicon pixels and silicon microstrips. An overview of the ID is shown in figure 2.7.



(a) Overview of the ATLAS ID.

(b) The different layers of the ID and their distance from the beam.

Figure 2.7: The ATLAS inner detector and its different layers [33].

Pixel detector

The pixel detector technology is based on the semi-conductor properties of silicon. A particle passing through the silicon pixels will liberate electrons in the material which creates electron-hole pairs. A bias voltage is applied across the p-n junction which is the boundary between p-doped (excess of holes and thus positively charged) and n-doped (excess of electrons and thus negatively charged) silicon. The voltage causes the electrons to flow to the positively charged silicon region and the holes to the negatively charged regions. The current that is thus produced is read out and registered as a particle hit.

As can be seen in figure 2.7, there are four silicon pixel layers in the pixel detector and four silicon microstrip layers in the Semi-Conductor Tracker (SCT). The pixel layers are the parts of the detector that are closest to the beam, with the closest layer only 3.3 cm away from the beam pipe. This closest layer is called the Insertable B-Layer (IBL) [34] and was added in the 2013–2015 extended technical stop before Run II of the LHC. This extra layer was added in order to maintain and improve the ATLAS tracking and b-tagging (see section 4.3.4) performance over Run II and beyond. Since the ID is so close to the beam pipe, the performance of the detector

will degrade over time due to radiation and this extra layer can offset the decrease in efficiency. It will also compensate for other failures in the pixel detector that inevitably happen over time. The IBL also helps dealing with readout inefficiencies associated with an increase in luminosity and improves the tracking precision because it adds an additional track measurement.

In addition to the four barrel layers, the pixel detector has three end-cap disks on each side. In total, the pixel detector has ~ 92 million readout channels (80 million excluding the IBL), which is more than half of the channels of the entire detector. The specifications of each part of the pixel detector are summarised in table 2.1. The high granularity of the pixel detector leads to a very high tracking precision. This high resolution is necessary to deal with the extreme concentration of particles around the interaction point, an example of which is given in figure 2.8.

Part of the pixel detector	**Total number of modules**	**Pixel size [μm^2]**	**Resolution [μm^2]**
Insertable B-Layer	224	50×250	$8(R\cdot\phi)40(z)$
Outer three barrel layers	1456	50×400	$10(R\cdot\phi)115(z)$
End-cap disks	288	50×400	$10(R\cdot\phi)115(R)$

Table 2.1: Specifications of the different pixel detector parts [31,34].

The pixel detector is essential for the identification of long-lived particles such as b-hadrons. The lifetime of a b-hadron is about 1.5×10^{-12} s which means that this particle travels a few mm in the detector before it decays. This results in a vertex of tracks displaced from the primary vertex. The pixel detector provides the precision needed for the identification of these *displaced vertices*, which is crucial in b-tagging (see section 4.3.4). The addition of the IBL has improved the b-tagging performance due to the increased vertex resolution it supplies.

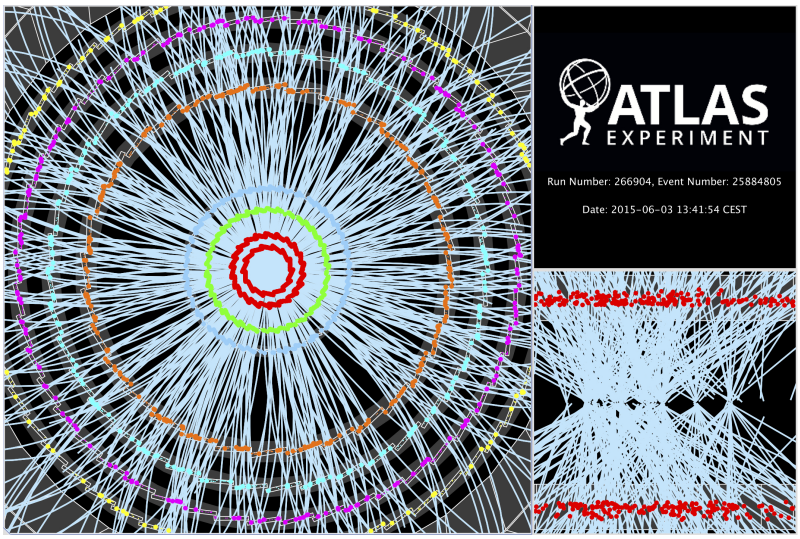


Figure 2.8: ATLAS event display from 2015 proton-proton collision, zoomed in on the ID [35]. The light blue lines indicate tracks reconstructed by the ID. The coloured dots represent hits in the silicon pixel layers (inner four rings) and silicon microstrip layers (outer four rings). There are 16 pile-up collision vertices in addition to the primary vertex in this event, some of which are resolvable in the bottom right view.

Semi-Conductor Tracker

The SCT is designed to contribute to the measurement of tracks, momentum, and vertex position. This subdetector of the ID relies on a detection system similar to that of the pixel detector. It consists of 4088 modules of silicon microstrips which are placed in four cylindrical layers in the barrel region and nine disk layers in each end-cap. Each layer has double-sided strip modules which are rotated by 40 mrad with respect to each other in order to allow for a two-dimensional particle hit. This provides a particle space point measurement in $R \cdot \phi$ and z for each layer traversed. The SCT provides a resolution of 17μm in $R \cdot \phi$ and 580μm in z (R) for the barrel (end-caps) [31]. It is thus less precise than the pixel detector (see table 2.1) but it covers a larger area which is important for tracks perpendicular to the beam.

Transition Radiation Tracker

Outside the silicon detectors is the Transition Radiation Tracker (TRT) which is designed to measure the curvature of the particle tracks rather than making a precise hit position measurement. The TRT consists of thin (4 mm diameter) straw tubes filled with a xenon-based gas mixture. The walls of the tubes act as cathodes and an anode wire is spun along the central axis of each tube. A charged particle passing through the tube will ionise the gas inside which leads free electrons to drift to the anode wire in the centre. The current recorded in the wire is then registered as a hit in the straw tube. Due to the straw layout parallel to the beam axis, the TRT has no sensitivity to the z direction of particles. However, it offers an accuracy of $\sim 130\mu$m in $R-\phi$ which is mainly determined by the drift time [31]. Although this accuracy is lower than the two silicon based systems of the ID, the TRT compliments the other systems because of the high multiplicity of hits: a track typically crosses about 36 straws in the barrel region.

The gaps between the straw tubes are filled with polymer material which creates transition radiation. This radiation is emitted when charged particles pass through an inhomogeneous medium or a material boundary and is stronger for high energy particles. In this way, the TRT allows for the distinction between light and heavy particles, especially between electrons and pions, and hereby helps in particle identification. The transition radiation photons are absorbed by xenon atoms in the gas mixture inside the straw tubes, which leads to a much higher current readout from the central anode wire. In order to distinguish between tracking hits and hits coming from transition radiation, the readout electronics use a separate low threshold (used for tracking) and a high threshold (used for transition radiation).

2.2.3 The calorimeters

The calorimeters are designed to precisely measure the energy and position of photons, electrons, and hadrons through the absorption of their energy in the calorimeter material. Muons, neutrinos, and other (hypothetical) weakly-interacting particles can punch through the calorimeter layers with ease. However, the calorimeters are designed to contain electromagnetic and

hadronic showers which means that electrons, photons, and hadrons should be stopped completely inside the calorimeter system. The calorimeter depth is therefore an important design factor. The ATLAS calorimeter system consists of the electromagnetic calorimeter (ECAL), the hadronic calorimeter (HCAL), and the forward calorimeters (FCAL)(see figure 2.9). The calorimeters cover the range $|\eta| < 4.9$ and use a variety of techniques for the requirements of different physics goals.

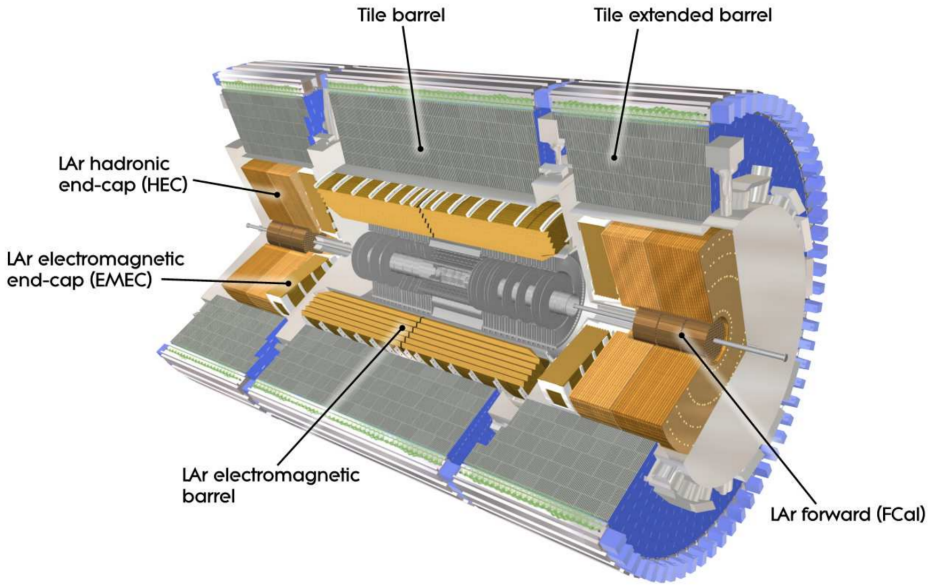


Figure 2.9: The ATLAS calorimeter system [33] showing the liquid argon (LAr) and scintillating tile components.

Electromagnetic calorimeter

The electromagnetic (EM) calorimeter is situated just outside the solenoid magnet surrounding the ID. It is composed of a barrel region covering the range $|\eta| < 1.475$ and two end-caps covering $1.375 < |\eta| < 3.2$. The junction between the barrel and end-cap, $1.375 \leq |\eta| < 1.52$, is called the *crack region*. This region is affected by additional material needed to service and cool the inner detector which leads to a reduction in performance due to layers of inactive material. This region is therefore excluded from most physics analyses requiring high precision electron or photon measurements.

Figure 2.10 shows the detailed structure of the ECAL. It employs stainless steel and lead as absorbing materials and liquid argon (LAr) as its active material, hence it is also called the *LAr calorimeter*. Charged particles traveling through the calorimeter ionise the liquid argon which produces electrons. An electric field is applied such that the electrons drift towards the readout electrodes. The LAr calorimeter has an accordeon geometry which allows for fast response and an absence of dead detector regions.

The granularity of the LAr calorimeter depends on the layer. The first layer is called the *presampler* and exists just of LAr without any absorber in front. This layer corrects for the energy losses in the ID and solenoid and has a resolution of $\Delta\eta \times \Delta\phi = 0.025 \times 0.1$ [31]. The second

layer is the first true sampling layer and has the highest precision with a barrel resolution of $\Delta\eta \times \Delta\phi = 0.0031 \times 0.1$ for $|\eta| < 1.40$ [31]. It is used to reconstruct the η position of particles. The third and fourth layers are coarser in η but finer in ϕ, with granularities of 0.025×0.025 and 0.050×0.025 respectively [31]. The third layer collects the largest part of the shower energy, whereas the fourth layer only collects the energy of the tail.

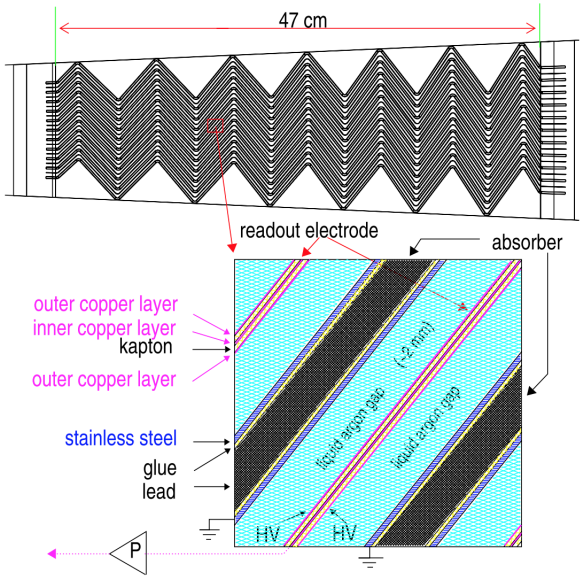


Figure 2.10: The ATLAS electromagnetic calorimeter showing its accordeon geometry [36].

Hadronic calorimeter

The hadronic (HAD) calorimeter, also called the *tile calorimeter*, consists of a barrel region ($|\eta < 1.7|$) and end-cap ($1.5 < |\eta| < 3.2$). The barrel uses alternating layers of steel as absorbing material and tiles of scintillating plastic as active material, providing a granularity of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ [31]. Hadronic showers cause the scintillating plastic tiles to emit light proportional to the deposited energy, which is then read out by wavelength shifting fibres into photomultiplier tubes. The detector components and readout system are presented in figure 2.11. The hadronic calorimeter end-caps use the same liquid argon technology as the electromagnetic calorimeter and provide a resolution in η, ϕ of 0.1×0.1 for $1.5 < |\eta| < 2.5$ and 0.2×0.2 for $2.5 < |\eta| < 3.2$ [31]. The hadronic calorimeter has a coarser resolution than its electromagnetic counterpart, but this is enough for jet reconstruction and E_T^{miss} measurements.

Forward calorimeter

The FCAL extends the calorimeter in the range $3.1 < |\eta| < 4.9$. It allows for the measurement of forward particle production and reduces the background radiation on the muon spectrometer. It is a LAr detector and is placed 4.7 m away from the interaction point on each side of the detector. The electromagnetic FCAL uses copper as absorbing material and the hadronic part uses tungsten.

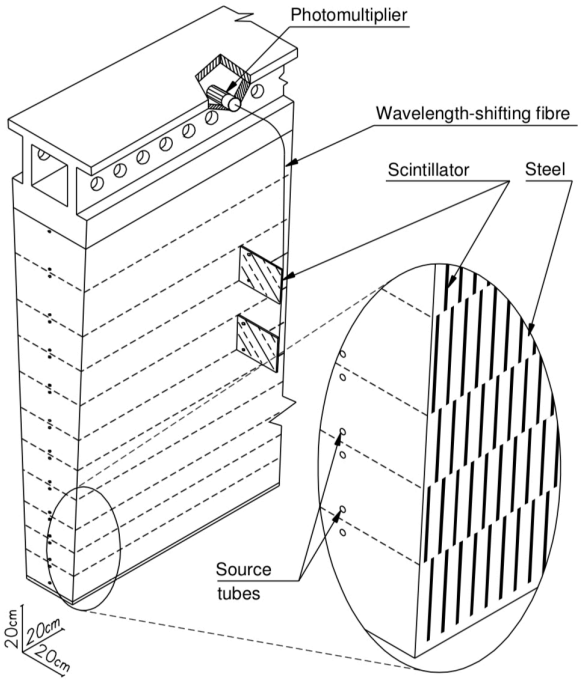


Figure 2.11: The ATLAS hadronic calorimeter showing its detector components and optical readout system [31].

Calorimeter response

The calorimeter response is measured as the ratio of the average calorimeter signal to the energy of the particle. The ATLAS calorimeters have a non-compensating response, which means that the calorimeter response to the EM shower component (e) and non-EM hadronic shower component (h) is not the same: $e/h > 1$. This is mainly the result of invisible energy present in the hadronic shower parts that does not contribute to the calorimeter signal [37]. This invisible energy comes from the release of nucleons from nuclei, neutrinos and highly energetic neutrons that escape the detector, and recoil energy. A local hadronic calibration is applied which provides some level of software compensation and aims to account for the invisible energy. This local calibration procedure is further described in section 4.1.

2.2.4 The muon spectrometer

Due to their long lifetime and low interaction rate, muons are in general not stopped by the calorimeter layers. Therefore, the outermost detector layer is the muon spectrometer which is designed to measure the muon momenta. This is accomplished by a toroidal superconducting magnet system which deflects the muons and high-precision tracking chambers which carry out the momentum measurement. The chambers are arranged in three concentric cylindrical shells around the beam axis for the barrel region. At the end-caps, the muon chambers are also three-layered and placed perpendicular to the beam. Figure 2.12 shows a cut-away view of the entire muon system with its different subdetector parts which are described below.

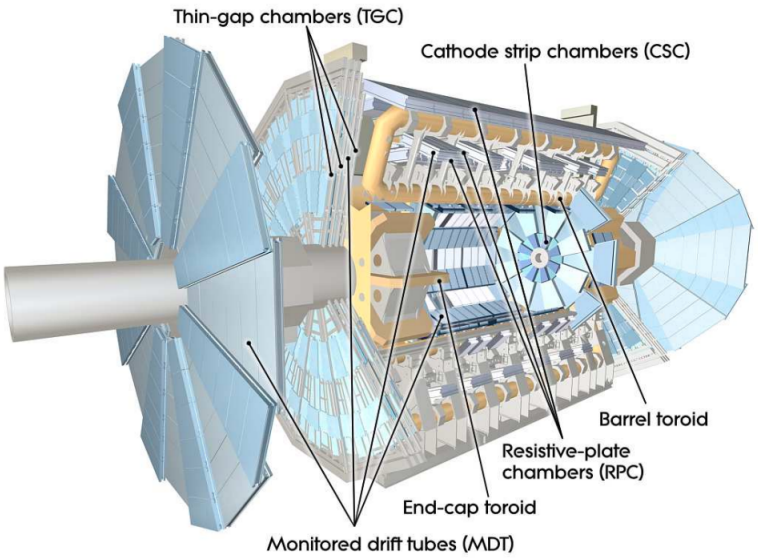


Figure 2.12: The ATLAS muon spectrometer showing its four subdetectors technologies [31].

Muon trigger system

The muon spectrometer has its own independent triggering system in the region $|\eta| < 2.4$, using Resistive Plate Chambers (RPC) in the barrel ($|\eta| < 1.05$) and Thin Gap Chambers (TGC) in the end-cap region ($1.05 < |\eta| < 2.4$). The RPC consists of two parallel resistive plates (an anode and a cathode) separated by a gas gap. Muons passing through the RPC ionise the gas which sets free electrons. These are accelerated by the electric field and start ionising more of the gas atoms which leads to a chain reaction of many accelerated electrons which is called an *electron avalanche*. These avalanches are read out by metallic strips on the outer surface of the plates. TGCs are multi-wire proportional chambers where an array of anode wires is placed between two graphite cathode layers and filled with a gas mixture. The wire-to-cathode distance (1.4 mm) is smaller than the wire-to-wire distance (1.8 mm). Again, a muon passing through the gas in these TGCs will ionise it and cause an electron avalanche which is collected on the nearest wire.

The muon trigger system selects interesting events containing muon candidates by providing identification of the individual bunch-crossings and measuring the muon track in the ϕ-plane which is orthogonal to the one measured by the tracking chambers. The RPCs have a spatial resolution of 1 cm and a very fast response time of about 1 ns which is necessary for the triggering. The resolution of the TGCs is slightly better at 5 mm and they have a response time of 4 ns.

Precision muon tracking

The muon trigger system defines Regions-of-Interest (RoI) in η and ϕ which are then scanned with precision by Monitored Drift Tubes (MDTs) in the region $|\eta| < 2.7$. These MDTs measure the muon's η coordinate and are made of aluminium tubes filled with an argon and carbon di-

oxide gas-mixture. A tungsten-rhenium wire at its centre produces a radial electric field. Muons ionise the gas and set free electrons that drift to the central readout wire. The drift time is the main limiting factor in the operation rate of the MDTs since it can reach up to 700 ns. Therefore, the MDTs cannot directly be used for triggering. They do provide a very high precision of about 80 μm per tube.

At large pseudorapidities ($2.0 < |\eta| < 2.7$), Cathode Strip Chambers (CSC) with higher resolution are used in conjunction with the MDTs. These help dealing with the large number of muons detected and the high background levels, since they have a better counting rate capability and time resolution. The CSCs are multi-wire proportional chambers where the cathode layers are segmented into strips in orthogonal directions (parallel to the wires and perpendicular to the wires) which allows for a 2D position measurement. They use the same gas-mixture as the MDTs and achieve a resolution of 60μm per plane in the η direction. The resolution achieved in the transverse plane is 5 mm.

2.3 Trigger and data acquisition

The data rate delivered to ATLAS by the LHC exceeds the recording and storing capabilities that are available. The main limiting factors are the ATLAS readout rate and worldwide data storage facilities of CERN. Therefore, we have a *trigger* system in place that selects the most interesting events to keep for physics analyses and decides which ones to delete. The ATLAS trigger system is divided into two levels, the first of which uses coarse data and works very quickly, whereas the latter uses more detailed information for the selection criteria of events and is slower. The first stage is hardware-based and is called the Level 1 (L1) trigger. The second stage is software-based and is referred to as the High Level Trigger (HLT).

A trigger menu is implemented with a list of trigger selections in operation at any point in the data-taking schedule. These menus define the exact L1 and HLT triggers to be used at a given luminosity. The trigger menu items can be pre-scaled, where e.g. a pre-scale of 10 means that only 1 in 10 events passing this particular trigger item are saved. This is chosen at random and allows for an optimal usage of the available bandwidth while the luminosity changes over an LHC run.

2.3.1 Level 1 trigger

The L1 trigger searches for large E_T, E_T^{miss}, and high-p_T muons, electrons, photons, jets, and τ-leptons. The muons are identified with the independent trigger system of the muon spectrometer. Coarse calorimeter information is used for the search for the other particles and large E_T and E_T^{miss}. The L1 trigger needs to make a decision in 2.5 μs to reduce the data rate from ~40 MHz to a maximum of 100 kHz. This stage of the trigger defines one or more RoIs in η, ϕ where the trigger selection process has identified potentially interesting objects. This RoI data contains information on the type of object identified, its coordinates, and its energy.

2.3.2 High Level Trigger

The RoIs defined by the L1 trigger are passed on to the next trigger stage: the HLT. This trigger is software-based and makes its event selection by using the detector data within the RoIs at full granularity and precision. Offline analysis algorithms are applied on fully reconstructed events in order to reconstruct the candidate physics objects. The use of tracking information allows for the identification of objects like electrons and muons. At this stage, the rate is reduced to approximately 1 kHz in ~ 250 ms.

2.3.3 Data acquisition

The DAQ system handles the actual moving, sorting, and saving of the data. When an event passes the L1 trigger, the RoI data gets temporarily stored in local memory. The HLT trigger subsequently requests this data for its own selection process. The events passing the HLT stage are transferred to permanent data storage at the CERN computer centres. Each event is assigned to a specific data stream indicating what type of analysis it can be used for. We distinguish between events good for physics analyses, detector monitoring and calibration, and a debug stream for events without a full trigger decision due to failures in the online system. Events in this last stream are studied offline and, if possible, are recovered and reprocessed after which they are added to their relevant data stream.

3 EVENT SIMULATION AND OBJECT RECONSTRUCTION

The prediction of the processes occurring in particle collisions at the LHC is a very important aspect of any ATLAS physics analysis. These predictions are needed for the design of an analysis as well as for the comparison of experimental data to the predictions from theory. The full calculation of the Matrix Element (ME) of the signal and background processes is usually only performed up to a few orders in perturbation theory. Therefore, we use the Monte Carlo (MC) event simulation technique which employs a factorisation method in which the events are divided up into more manageable parts. The ME can be computed up to a fixed order in perturbation theory, while the description of the parton shower (PS) and final state can be done with phenomenological models. The MC technique relies on the repeated (pseudo)random sampling of variables from probability distributions in order to obtain numerical results. The process of event simulation using MC methods is explained in section 3.1.

In order to interpret the experimental data as well as the simulated MC data, the ATLAS collaboration has developed a software framework which reconstructs physics objects in the events from tracks and energy deposits recorded in the detector. The reconstruction uses information from different parts of the detector, depending on the object to be constructed. The object reconstruction process is described in section 3.2.

3.1 Monte Carlo simulation data

We use simulated collision events generated with the MC method to simulate our signal process, possible background processes, and pile-up, in order to predict event rates and topologies. In order to make a direct comparison between *pp* collision data and simulated data, the MC events are fed through a simulation which models the detector response to stable particles.

3.1.1 Event generation

The generation of MC events happens in several stages. The energy at which the calculation is split is determined by the *factorisation scale*, μ_F, which separates the long-distance soft processes from the short-distance hard interaction. This means that the partonic processes are

described by the ME above the factorisation scale, and by the PS and Parton Distribution Functions (PDFs) below this scale. A schematic of a proton-proton collision involving all the steps in the event simulation is shown in figure 3.1.

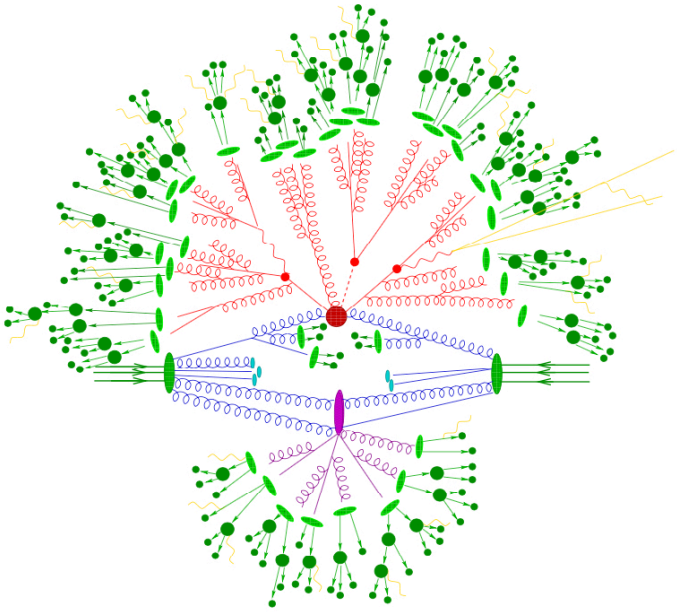


Figure 3.1: Schematic of a proton-proton collision in the LHC containing all steps of the event generation [38]. The protons are represented by the two large green ovals and their partons by the blue lines. One parton of each of the two protons collide with each other to form the hard scatter of the event, shown as the red circle in the middle. The hard scatter is surrounded by a red tree-like structure representing the parton shower. The purple structure at the bottom signifies a secondary scatter in the event which is referred to as underlying event. The hadronisation of partons into colourless hadrons is indicated by light green ovals, whereas the dark green structures indicate the hadron decays. The wavy yellow lines indicate soft photon radiation and the straight yellow lines represent leptons and neutrinos.

Parton distribution functions

We cannot explicitly compute the flavour and momentum of partons coming from the protons in collisions since this is a non-perturbative process. Therefore, we use PDFs to describe the momentum distribution and flavour of the partons. These PDFs are extracted from experimental measurements of fixed-target and collider experiments. They give the probability that a parton of a specific type is found inside a proton carrying a certain fraction of the proton momentum. These PDFs do not depend on the process under consideration and are thus universal. The PDF information is used in the MC simulation for the ME, PS, and UE.

The PDFs can be determined to different levels of accuracy, usually at leading order (LO) or next-to-leading order (NLO), and sometimes at next-to-next-to-leading order (NNLO) in QCD. The PDF also makes a choice on the flavour scheme used in the calculation. The most common distinction is made between the four-flavour (4F) and five-flavour (5F) schemes. The former implements massive b-quarks whereas in the latter scheme the b-quarks are treated the same as other massless partons. In the 5F scheme, the b-quarks are included in the initial state, i.e. in the partons which can be found within the proton. In the 4F scheme, the partons inside the

proton are limited to gluons and the four lightest quarks, and the b-quarks are included in the final state.

There are many different PDF sets that are used for the calculation of MC simulated events. Two of these are used in the $t\bar{t}H$ analysis discussed in this thesis: NNPDF [39] and CT10 (CTEQ) [40–42].

Hard scatter

The hard scatter is also called the ME and refers to the computation of the Feynman diagrams of the process of interest up to a fixed order in perturbation theory. Most processes are now calculated to NLO precision in QCD. This fixed-order calculation introduces a dependence on the *renormalisation scale*, μ_R, which originates from the renormalisation procedure. This procedure cancels out ultraviolet (UV) divergences in higher-order computations. The renormalisation scale is the value at which the running coupling is evaluated and is usually set equal to the factorisation scale, μ_F.

The hard scatter is computed at the highest energy scales and describes how the partons inside the protons interact and produce outgoing particles, as shown by the red central circle in figure 3.1. The momenta of the initial state partons are randomly sampled from the PDF. The final stages of the ME calculation can overlap with the start of the calculation of the next step carried out by the PS generator. In order to remove this overlap, a *matching procedure* is defined which determines the separation between the phase spaces covered by the ME and by the PS.

The hard scatter MC generators rely purely on theoretical predictions and the numerical computation of the process under consideration. The generators used in the $t\bar{t}H$ analysis are MadGraph5_aMC@NLO [43], Powheg-Box [44, 45] and Sherpa [46]. The first generator computes the ME at LO or NLO and matches it to the PS with the MC@NLO method [47]. The second generator computes the ME to NLO and uses the Powheg method [44, 48] for matching ME to PS. Sherpa is a LO/NLO generator containing the ME calculation as well as a PS algorithm and therefore does not need a matching scheme. It can be interfaced to additional libraries to compute loop amplitudes, such as via OpenLoops [49].

Parton shower

The PS step provides corrections to the ME calculation as it describes the evolution of partons from the hard scatter, both in the initial and final state. These corrections account for the evolution in momentum transfer from the high scales of the hard scatter to the low scales of hadronisation (about 1 GeV). The PS is represented in figure 3.1 as the red tree-like structure surrounding the hard scatter. In this step, the effect of higher order calculations, which were not included in the ME, can be taken into account. However, these corrections cannot be calculated exactly and therefore an approximation scheme is used in which only the dominant

contributions in each order are included. These are extra emissions via QCD or QED processes, including soft and collinear emissions.

The shower generators rely on theoretical predictions tuned to data. The three generators used for modelling the PS in the $t\bar{t}H$ analysis are PYTHIA [50, 51], HERWIG [52, 53], and SHERPA [46]. The first two are interfaced to any of the ME generators mentioned above, whereas SHERPA has its own ME calculation as well as showering and hadronisation models. PYTHIA orders the emissions in the PS by transverse momentum (see e.g. reference [54] for details) whereas HERWIG orders them by opening angle (see e.g. reference [38] for details).

Hadronisation

This final step describes the hadronisation of the partons into colour neutral particles which occurs due to colour confinement in QCD. It also includes the decay of unstable hadrons to stable final-state particles. These processes occur at the cut-off scale for the PS at around 1 GeV. The hadronisation and consecutive decays are shown in green in figure 3.1. Since hadronisation is also a non-perturbative process, it is based on empirical models tuned to data. There are two commonly used models: the Lund string model [55] and the cluster model [56]. The former model is used by PYTHIA and transforms the partons into colour neutral particles without intermediate states. The latter model is used by HERWIG and SHERPA and forms the colour neutral particles from partons via intermediary clusters of objects.

Underlying event

The UE refers to the interaction of partons which are not involved in the hard scatter of the event, as shown in purple in figure 3.1. These interactions can lead to soft additional jets in the event. Due to the low energy of these processes, their phenomenological models depend on free parameters which are tuned to data [57].

3.1.2 Detector simulation

The output of the simulation described above is also called the truth level and represents the physics objects in the event before they have interacted with the detector. The truth level events are made up of four-vectors of all 'stable' particles (typically mean lifetimes of $\tau \geq 3 \times 10^{-11}$ s, i.e. $c\tau \geq 9$ mm) produced in the event after hadronisation and hadron decay. These events include all necessary kinematic information but still need to go through a detector simulation in order to be compared to data.

The detector simulation first replicates the interactions of the particles with the various components of the detector, after which it converts the tracks and energy deposits to electronic signals. This results in a simulation sample with the exact same format as the experimental data. The detector simulation usually includes a very detailed description of all the subdetectors and is called full simulation, or *fullsim* for short. Since we typically need to produce several

million events per sample and the fullsim takes several minutes per event, a faster detector simulation is also available: the fast simulation (*fastsim*). In ATLAS, the fullsim is generated with the GEANT4 software [58] and the fastsim with ATLAS FAST-II (AF2) [59].

AF2 reduces the overall simulation time by about a factor of ten and is needed in cases where, due to time or computing resource constraints, the fullsim cannot be run for all samples. The AF2 simulation reduces the resources necessary for computation because it simplifies the showering of particles in the calorimeter which takes up about 90% of the computing time. It uses the full GEANT4 simulation for the ID and muon spectrometer, whereas the FastCaloSim simulation [60] is used for the calorimeters. This FastCaloSim uses parametrisations of the calorimeter response to photons, electrons, and pions (used for all hadrons) to directly deposit the energy of single particle showers. The AF2 simulation is useful in some analyses, but any analysis that needs detailed calorimeter information (e.g. involving jet substructure, see section 4.4.1) will suffer from the inherently less accurate parametrisations used by the FastCaloSim.

In the $t\bar{t}H(H \to b\bar{b})$ analysis presented in this thesis, checks are made in order to determine whether the fullsim is needed or if the fastsim suffices. Some examples of these studies are shown in figure 3.2. This figure compares the GEANT4 fullsim with AF2 fastsim detector simulation for the $t\bar{t}$ background sample only. All other samples (these will be detailed in section 5.4) are generated with the full detector simulation. These studies are first carried out for a loose boosted $t\bar{t}H$ selection in which we require one large trimmed jet which is top-tagged (see section 4.4 for more details) and three small jets which are tagged as containing a b-hadron (see section 4.3.4). Two of the studied variables in this selection are shown in figures 3.2(a) and (b); these are two jet substructure variables which will be explained in section 4.4.1. We see a larger discrepancy between data and MC for the AF2 simulation than the full GEANT4 simulation. Especially in the τ_{32} variable, a very clear slope can be observed in the data/MC comparison when using the fastsim. Figures 3.2(c) and (d) show the data-MC comparisons for two variables in the final boosted signal region used for the analysis presented in this thesis. This signal region requires two large reclustered jets and five small jets of which at least four are b-tagged (see section 5.5 for more details). Again, we see an increase in incompatibility between data and MC when we use the fastsim instead of the fullsim. For this reason, we use the full GEANT4 detector simulation in our analysis and we use the AF2 fast simulation only for the alternative samples used to evaluate the systematic uncertainties (see section 7.3).

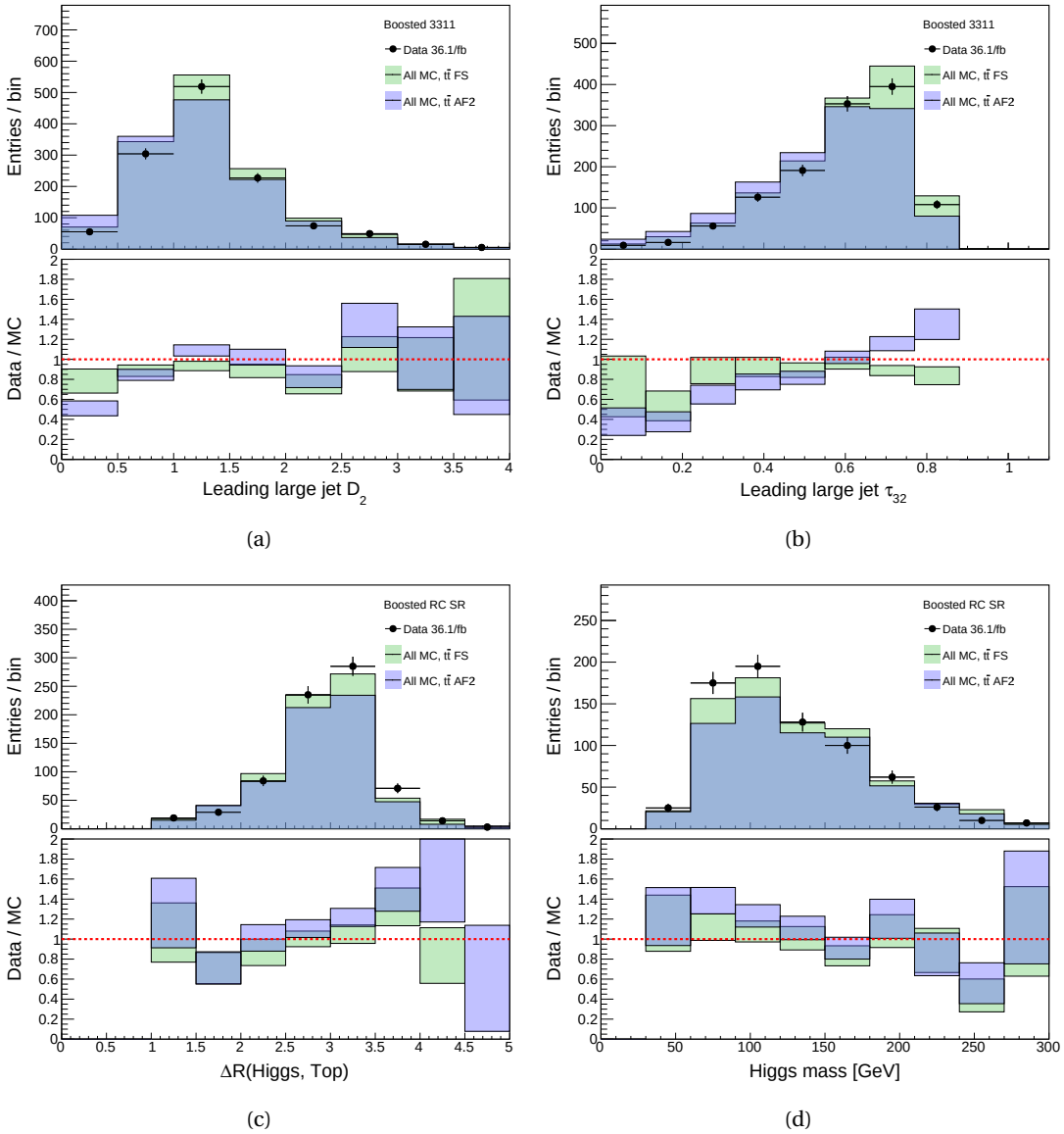


Figure 3.2: Comparison between the GEANT4 fullsim and AF2 fastsim detector simulation for the $t\bar{t}$ MC sample, compared to 36.1 fb^{-1} of ATLAS data from 2015–2016. The bars in the ratio plots represent the total statistical uncertainty. Figures (a) and (b) show variables in the loose boosted $t\bar{t}H$ region requiring one large top-tagged jet and three small b-tagged jets. Figures (c) and (d) show two variables in the boosted $t\bar{t}H$ signal region (see section 5.6.1).

3.2 Object reconstruction

The data recorded by the ATLAS detector need processing in order to identify physics objects in the events, such as electrons and muons. In the ATLAS collaboration, this reconstruction is done with the Athena software framework [61]. The physics objects are defined by a set of loose requirements which can later be tightened depending on the physics analysis. The object reconstruction is performed using information from different subdetectors depending on the object under consideration.

An overview of the particle identification in ATLAS is shown in figure 3.3. All charged particles leave hits in the ID and are deflected by the magnetic field applied here (see section 2.2.1). The direction and amount of track deflection gives us information about the momentum of the particle. Electrons and photons are mostly stopped by the ECAL; the electron leaves a track in the ID whereas the photon does not. Protons leave tracks in the ID as well and are stopped in the HCAL together with neutrons. However, since neutrons have no charge they do not create hits in the ID. Both muons and neutrinos pass through the entire detector, but muons leave tracks and energy deposits whereas neutrinos are invisible to the detector. We infer the presence of neutrinos by using the missing transverse energy (MET) (see section 3.2.5).

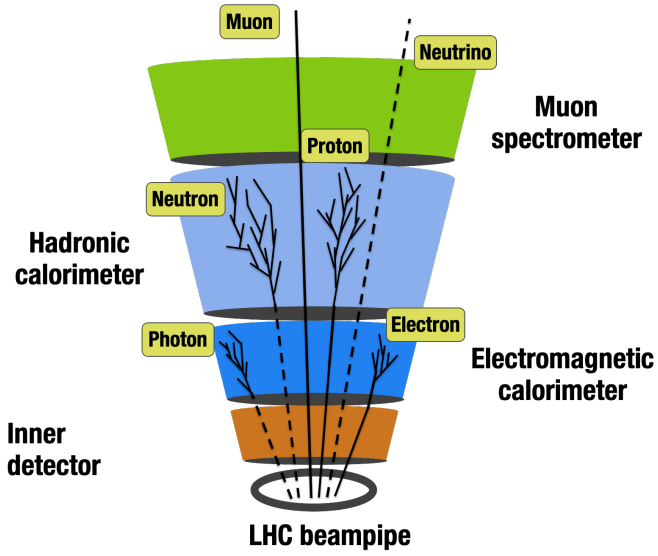


Figure 3.3: An overview of particle identification in the ATLAS detector. The solid lines indicate tracks and energy deposits left by charged particles. The dashed lines represent neutral particles that are invisible to the detector until they are stopped in the calorimeters and produce a particle shower. The neutrinos pass through the entire detector without leaving a signal.

3.2.1 Tracks and vertices

The tracking [62, 63] and vertexing [64] algorithms are both based on information from the ID (see section 2.2.2). The tracks are essential for electron, muon, and vertex reconstruction. In turn, the vertices are crucial for flavour tagging. A charged particle passing through the ID generates hits in the various layers which are combined to form the particle track. This

combination is done using an *inside-out* pattern recognition algorithm [62] which starts from seeds in the innermost layer of the ID and works outwards through the pixel, SCT, and TRT layers. An *outside-in* algorithm known as back-tracking is used to take into account any hits that were not selected by the inside-out algorithm. The back-tracking is seeded in the TRT and works inwards towards the pixel layers.

The reconstruction of primary vertices is based on the grouping of reconstructed tracks with an adaptive vertex fitting algorithm [65]. In order to be used in the construction of a vertex, the track must have $p_T > 400$ MeV and $|\eta| < 2.5$. The primary vertices are required to have at least two tracks associated to them and to lie inside the *beamspot* area. This is the region around the interaction point where the two proton beams overlap. The vertex reconstruction is an iterative process and terminates when either all tracks are associated to a vertex or no additional vertices can be constructed.

The main, hard scatter, vertex of an event is defined as the vertex with the highest sum of associated track p_T. The other primary vertices constitute the pile-up. We can also identify secondary vertices which are incompatible with the beamspot region. These are formed by tracks displaced from the primary vertex which can be created by particles with a long lifetime such that their decay length is significantly large (a few mm), such as b-hadrons. This feature is used in the identification of jets containing such particles, as described in section 4.3.4.

3.2.2 Electrons and photons

Electrons and photons are reconstructed in the central detector region within $|\eta| < 2.47$ [66–68]. The first step in the reconstruction is the clustering of calorimeter energy deposits from seeds. These seeds are energy deposits with $p_T > 2.5$ GeV, where the energy of all calorimeter layers is added together. This threshold was chosen to optimise the reconstruction efficiency while minimising the contribution of noise from electronic or pile-up sources. The clustering is done with a *sliding-window* algorithm which looks for clusters by performing a scan over blocks of size 3×5 in units of $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$. This corresponds to the granularity of the middle sampling layer of the ECAL, in which most of the energy of the electrons is deposited.

When the energy clusters are identified, they are matched to tracks in the ID. Tracks with $p_T > 500$ MeV are extrapolated from their last measured point in the ID to the middle sampling layer of the ECAL. The position in η, ϕ space extracted from this extrapolation is compared to the position of a seed cluster in the ECAL layer and the two are matched if their separation is $|\Delta\eta| < 0.05$ and $|\Delta\phi| < 0.1$. The clusters of energy are labelled as electron candidate when at least one track is matched to the seed cluster, and as a photon candidate when no tracks are matched. If multiple tracks are matched to a cluster, a primary track needs to be identified in order to determine the electron kinematics and charge. Tracks that have hits in the pixel or SCT subdetectors are given priority, and the one closest to the centre of the cluster is picked.

If a successful match is made between an energy cluster and a track, the cluster size is enlarged to 3×7 units in the barrel region ($|\eta| < 1.475$) and 5×5 units in the end-caps ($1.375 < |\eta| < 3.2$).

These larger clusters are then used to determine the electron candidate energy.

After the electron and photon reconstruction, identification algorithms are applied in order to rule out potential misidentifications. For example, hadron jets or converted photons (a $\gamma \rightarrow e^+e^-$ conversion due to interaction with the detector material) can be mistaken as true electrons. These objects mimicking a lepton are referred to as *fakes*. In order to separate the true electrons from the fakes, a likelihood discriminant is used which incorporates tracking, calorimeter, and combined track-cluster variables.

Three working points (WPs) are defined for the identification of electrons: *loose*, *medium*, and *tight*. These provide increasingly improved background rejection with, consequently, a lower identification efficiency. The tight WP thus provides the highest purity of real electrons to fakes. The improved background rejection is obtained by tightening the requirements on the variables at each step. The signal efficiency to identify electrons is given by the ratio of the number of electrons passing the identification requirement to the total number of electron candidates. The efficiencies for electrons with $E_T = 25$ GeV from simulated $Z \rightarrow ee$ decays range from 78% for the *tight* WP to 90% for the *loose* WP, and increase with E_T. The $t\bar{t}H$ analysis discussed in chapter 5 uses electrons passing the *tight* WP. For photons, a *loose* and a *tight* WP are defined (see reference [69] for more details).

In order to further suppress the fakes contribution, isolation requirements are defined for the electrons [68]. These requirements are based on track and calorimeter variables which quantify the energy of the particles around the electron candidate in a cone of $\Delta R \leq 0.2$. In the $t\bar{t}H$ analysis discussed in this thesis, the *gradient* isolation operating point is used. This is implemented as a sliding cut on the p_T of tracks and E_T of cluster deposits and becomes more stringent as the electron E_T decreases.

3.2.3 Muons

The muon reconstruction uses tracks from the ID and the muon spectrometer [70]. There are four muon types which are defined by different identification criteria: *combined*, *segment-tagged*, *calorimeter-tagged*, and *extrapolated*. The analysis in this thesis uses combined muons for which track reconstruction is first performed independently in the ID and the muon spectrometer, after which a combined track is formed with a global fit. Most muons are reconstructed with an *outside-in* approach where they are first reconstructed in the muon spectrometer and then extrapolated inwards to match a track in the ID. A small fraction of muons, about 0.5% for the muons used in the $t\bar{t}H$ analysis, are reconstructed with an *inside-out* approach where ID tracks are extrapolated outwards to match the muon spectrometer tracks.

Just like for the electrons, identification criteria are applied to the muon candidates in order to suppress background, which mainly comes from pion and kaon decays. These criteria are based on the χ^2 of the combined track fit, track quality requirements, and variables related to the difference in charge and p_T measured for the muon candidate in the ID and muon spectrometer. Four WPs are defined for muon identification: *loose*, *medium*, *tight*, and *high-p_T*. In

the analysis presented in this thesis, the *medium* WP is used for muon reconstruction; this is the ATLAS default.

In order to identify the muons coming from heavy particle decays such as the W boson, we apply isolation criteria to the muon candidates. The muon isolation works similar to the electron isolation: one track-based and one calorimeter-based isolation variable are employed in order to measure the amount of energy surrounding the muon candidate in a cone of $\Delta R \leq 0.3$. Seven isolation WPs are provided; we use the *gradient* isolation criteria in the $t\bar{t}H$ analysis. This leads to tighter isolation requirements for muons with lower p_T.

3.2.4 Taus

The tau leptons are treated as electrons or muons when they decay leptonically but are reconstructed as taus (τ_{had}) when they decay hadronically ($\sim$ 65% of the time). The τ_{had} are separated from hadronic jets by the use of a Boosted Decision Tree (BDT) (see section 5.7.1) which incorporates information from the tracks and energy clusters of the jets [71]. The tau leptons are not explicitly used in the $t\bar{t}H(H \to b\bar{b})$ analysis, but they are reconstructed in order to establish orthogonality with other $t\bar{t}H$ searches. This is needed for the final combination which will be discussed in section 8.3. A veto is applied to events if they contain one (two) or more τ_{had} lepton candidates in the dilepton (single-lepton) channel of the $t\bar{t}H(H \to b\bar{b})$ analysis.

3.2.5 Missing transverse energy

The MET, also labelled E_T^{miss}, refers to an imbalance in the visible transverse momentum in an event. It is based on the principle of conservation of energy and momentum and is used to infer the presence of particles that are invisible to the ATLAS detector, such as neutrinos. The missing energy is measured in the transverse plane (x, y) only; a full missing energy measurement would be impossible because the energy of the initial partons is unknown. However, we can assume that the initial partons have negligible momentum in the transverse plane because *pp* collisions happen along the z-axis. The MET is defined as the negative vector sum of the momenta of all other reconstructed objects in the event [72]. An additional soft term is included as a correction for detector signals which were not associated to any reconstructed object, such as tracks in the ID or energy deposits in the calorimeter.

3.2.6 Jets

Jets are collimated sprays of hadrons which are visible in the ATLAS detector. They serve as a proxy for the quarks and gluons involved in the *pp* collisions. Jets are constructed with a small or a large radius in order to catch all the energy deposits and/or tracks originating from one object together into one jet. Since jets play a crucial role in the boosted $t\bar{t}H(H \to b\bar{b})$ analysis, chapter 4 is dedicated to their definition, reconstruction, and calibration.

4 JETS

The analysis of data coming from the high-energy LHC collisions aims to study the quarks and gluons created in this process. However, we cannot look at these particles directly because they hadronise almost instantly after being produced due to colour confinement. This leads to a collimated spray of energetic hadrons which are visible in the ATLAS calorimeters as clustered energy deposits: we call these *jets*. We attempt to group inputs from common sources together into a single jet such that we can use these jets as a proxy for the original partons.

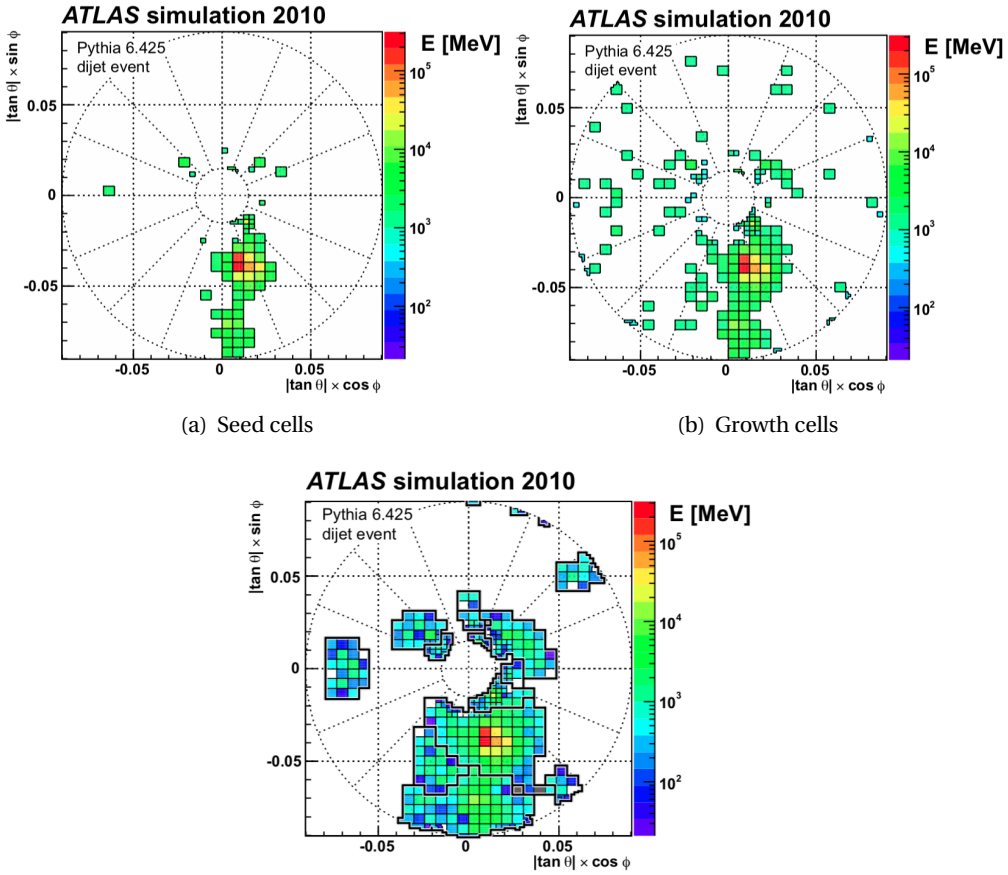
Using jets in a consistent way requires some prescription on how to define them. In this chapter we will describe how this is done in ATLAS. First we need to define what we use as inputs to our jets (section 4.1) and then we will look at *jet algorithms*: sets of rules that define how to group these inputs together into a jet (section 4.2). After this we will discuss the most common definitions of jets used in ATLAS analyses and how to calibrate them (section 4.3). We will then move on to some studies concerning different methods of defining large jets (section 4.4). The definitions, methods, and studies described in this chapter set the scene for the $t\bar{t}H$ analysis discussed in chapter 5.

4.1 Jet inputs

In ATLAS, we use energy deposits left in the calorimeter and tracks measured by the ID as inputs to our jet reconstruction algorithms. Before the calorimeter energy deposits are put into the jet clustering algorithms, a topological clustering of the calorimeter cells is carried out [73]. Because of imperfections in the detector system and the influence of pile-up, each calorimeter cell is expected to record a base level of noise. The average noise is estimated for each run year according to:

$$\sigma_{\text{noise}} = \sqrt{(\sigma_{\text{noise}}^{\text{electronic}})^2 + (\sigma_{\text{noise}}^{\text{pile-up}})^2} \tag{4.1}$$

where $\sigma_{\text{noise}}^{\text{electronic}}$ is the electronic noise and $\sigma_{\text{noise}}^{\text{pile-up}}$ is the noise from pile-up. As described in section 2.1.2, pile-up is proportional to the instantaneous luminosity which decreases over the course of an LHC fill. Therefore, the average calorimeter cell noise from pile-up is expected to decrease over the course of an LHC fill as well. Note, however, that the average noise definition



(a) Seed cells

(b) Growth cells

(c) Final topoclusters including the boundary cells

Figure 4.1: The three stages of defining topoclusters in the ATLAS calorimeter for a simulated dijet event. In (a) we see the cells that seed the topoclusters, in (b) the cells controlling topocluster growth, and in (c) the boundary cells which make up the final topocluster [73].

needs to be fixed before data-taking in order to reconstruct the data efficiently. This means that the definition is not optimal for most of the pile-up spectrum throughout an LHC fill.

In order to select the significant signal from the background noise, we need to define an energy threshold. The first step in the topological clustering is therefore to define *seed cells* which are required to have an energy of at least 4 times the average expected noise in that cell ($|E_{\text{cell}}| > 4 \times \sigma_{\text{noise}}$), see figure 4.1(a). These seed cells form the first stage proto-cluster. We then define a group of *growth cells* which neighbour the seed cells and have an energy of at least 2 times the average expected noise in that cell ($|E_{\text{cell}}| > 2 \times \sigma_{\text{noise}}$), see figure 4.1(b). Neighbouring here means that two calorimeter cells are directly adjacent in the same layer or across layers. If any of the neighbour cells happens to be an original seed cell, the two proto-clusters are merged. If a growth cell neighbours two separate proto-clusters, they are all merged. Finally, any cells directly neighbouring the growth cells (called *boundary cells*) are taken into the proto-cluster as well, with no requirement on the energy intensity ($|E_{\text{cell}}| > 0$). This step then defines the final three-dimensional topocluster as shown in figure 4.1(c).

Before the topoclusters are used in the reconstruction of jets, they are calibrated to either the electromagnetic or hadronic scale response. The latter calibration accounts for the non-

compensating calorimeter response (see section 2.2.3) as well as for signal losses due to the topological clustering algorithm and energy deposited in inactive detector material. The EM calibration calibrates the calorimeter cells' signal and its average expected noise to the response from electrons. This means that the topocluster accurately reconstructs the energy deposits of electrons and photons but it does not attempt to improve the non-compensating calorimeter behaviour or above-mentioned signal losses. The topocluster mass is set to zero in both calibrations which means that $E = p$.

The hadronic calibration is called the local hadronic cell weighting (LCW) calibration since it includes cell signal weighting. The first step of this calibration procedure is the classification of topoclusters as HAD or EM in origin, since the calibrations and corrections applied are dependent on the type of energy deposit. Hadronic showers tend to be less dense and penetrate deeper into the calorimeter than electromagnetic showers. Therefore, the classification of topoclusters is based on the energy of the cluster (signal energy density ρ_{cell}) and its position (longitudinal depth λ_{clus}), as shown in figure 4.2. These two variables define a dynamic scale from which the topoclusters get a specific calibration and correction, depending on the probability that the topocluster was generated by an EM or HAD shower. The main difference between the calibrations and corrections is that they are significantly smaller for the EM-like topoclusters than for the HAD-like topoclusters.

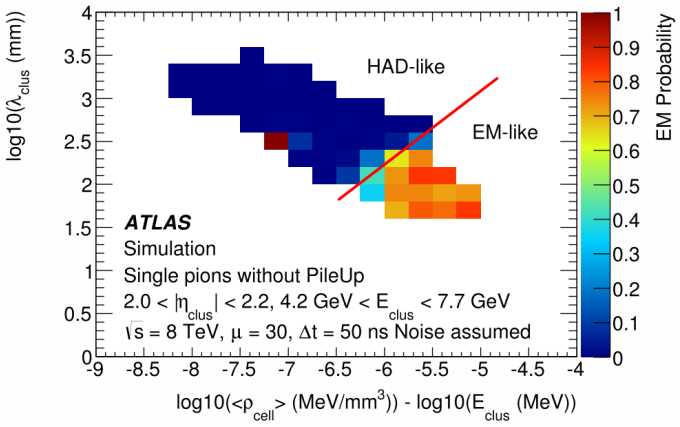


Figure 4.2: The classification of topoclusters as hadronic or electromagnetic in origin based on the signal density ρ_{cell} and longitudinal depth λ_{clus} [73].

Each topocluster should represent one incident particle, but if two or more particles were produced very close together they might merge into one topocluster. In this case, we need the tracking information from the ID to distinguish between the incident particles. Tracking information can also be useful for rejecting pile-up jets, as will be described in section 4.3.

4.2 Jet algorithms

Once we have the energy constituents of our jets, the topoclusters, we need to define a recipe that tells us which topocluster belongs to which jet. This recipe is called a *jet algorithm* and there is no unique way to do this. The first ever jet algorithm was developed in the 70's [74]

and was a top-down cone algorithm. Drawing a cone around a collimated spray of particles is an intuitive jet definition. However, these cone algorithms struggle with infrared and collinear (IRC) safety. For an event to be IRC safe, the set of hard jets found in the event should stay the same when a collinear splitting or a soft emission is added to the event. Collinear safety implies that the jet boundary is not affected if a single particle is replaced by two collinear particles of half the original energy. Infrared safety implies that the jet clustering is driven by the hardest energy deposits and ignores the low energy coming from soft radiation of the initial parton.

In order to circumvent the IRC unsafety of cone algorithms, we now use sequential recombination algorithms which have a bottom-up approach. These algorithms iteratively combine the closest sets of particles, sometimes dependent on their transverse momentum. There are three algorithms that are most widely used in hadron collider experiments. They can all be defined according to the following equations:

$$d_{ij} = \min(p_{T,i}^{2p}, p_{T,j}^{2p}) \frac{\Delta R_{ij}^2}{R^2}, \tag{4.2}$$

$$\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \tag{4.3}$$

$$d_{iB} = p_{T,i}^{2p}. \tag{4.4}$$

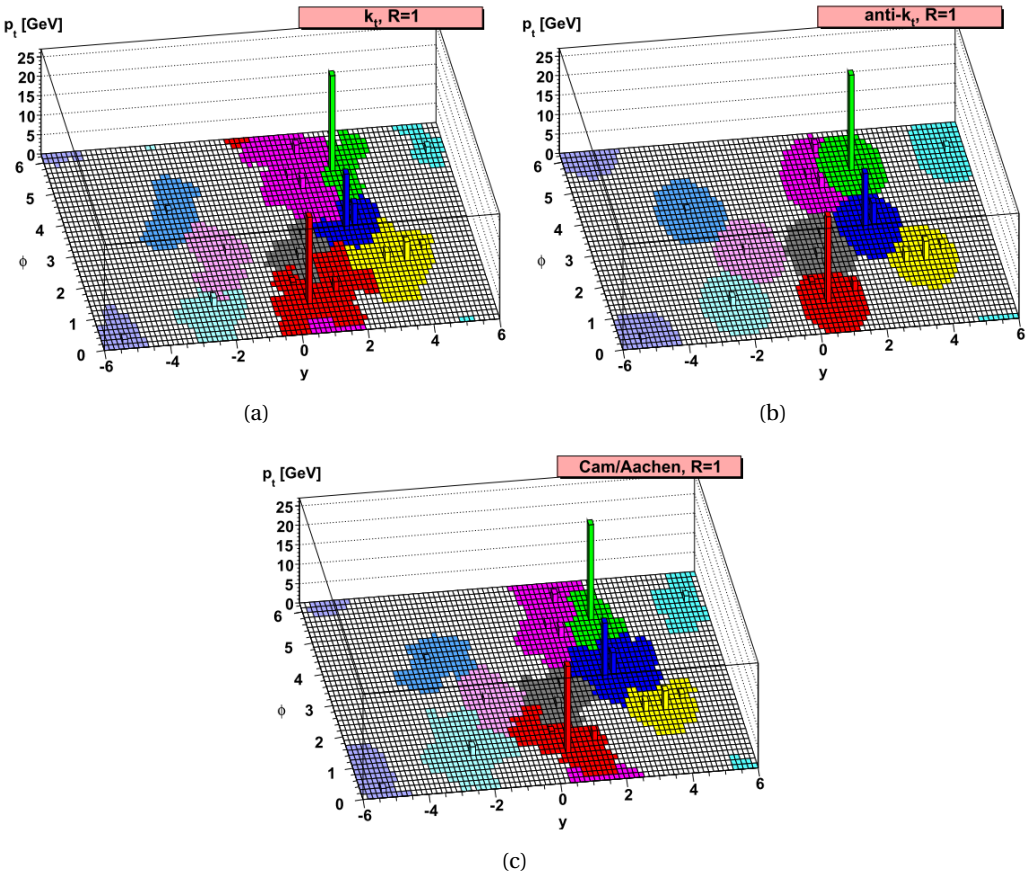
Equation 4.2 describes the distance parameter d_{ij} between two particles i and j in the event, with equation 4.3 defining the R parameter (y is the rapidity and ϕ the azimuthal angle, see section 2.2). The distance between the particle i and the hadron beam, d_{iB}, is described in equation 4.4. The parameter p in both equations 4.2 and 4.4 determines which of the three jet algorithms we are considering. Note that d_{ij} and d_{iB} are invariant under longitudinal boosts.

***k_T* algorithm:** The k_T algorithm [75, 76] clusters together close and low-momentum (soft) particles first. For this algorithm, the p parameter mentioned above is set to 1. For each pair of particles i, j in an event, we compute the distance parameter as given by equation 4.2. We then select the smallest value, $d_{\min}$, of all d_{ij} and compute the particle-beam distance for the particle i, as given by equation 4.4. If $d_{ij} < d_{iB}$, the algorithm recombines particles i and j into a single new particle and starts the process again by redoing the computation of all distance parameters and particle-beam distance as defined above. However, if $d_{iB} < d_{ij}$, the particle i is declared a final-state jet and removed from the list of particles, after which the distances above are calculated again. This process is repeated until there are no more particles left. The parameter R in equation 4.2 determines the size of the jet since d_{iB} will be smaller than d_{ij} for any j if the particle i has no other particles close to it within the distance R.

Anti-*k_T* algorithm: The p parameter is set to -1 for this algorithm and the same procedure is carried out as for the k_T algorithm. This leads to the anti-k_T algorithm [77] clustering together close and high-momentum (hard) particles first. The jets grow outwards around a centre of high-p_T topoclusters which results in circular cone-shaped jets with size R.

Cambridge/Aachen algorithm: Also called the C/A algorithm [78, 79], this jet definition uses a distance measure only based on the geometrical scale and is thus independent of the particles' energy and momentum. For this algorithm, we set the p parameter to 0 which leads to a check at every step of the algorithm whether $d_{ij} = \Delta R_{ij}^2 / R^2$ is larger or smaller than 1. It thus starts by clustering together the pair of particles closest to each other into one object, and repeats this process until all objects (the final-state jets) are separated by $\Delta R_{ij} > R$, where R is the size of the jet.

The three jet algorithms each produce slightly different jets with different catchment areas and boundaries, as shown in figure 4.3. The anti-k_T algorithm is favoured because it produces circular jets. The k_T and C/A algorithms produce geometrically irregular jets which complicates detector corrections and corrections from non-perturbative sources. This is due to the fact that irregular jet boundaries mean that they can extent beyond a distance R from the jet momentum. This makes it difficult to define a detector region in which all jets are fully contained.



(a) (b) (c)

Figure 4.3: A simulated event clustered with the three different sequential recombination jet algorithms, showing the shape of the jet boundaries [77].

Besides the fact that the anti-k_T algorithm produces circular jets, it has a range of other benefits over the other recombination algorithms. The performance of the different jet algorithms was tested in reference [80] where it was shown that the anti-k_T algorithm exhibits the best jet reconstruction efficiency, trigger matching performance, and stability under pile-up. It is also one of the fastest algorithms and requires a low memory consumption. It is therefore the stand-

ard algorithm used to define jets in ATLAS analyses.

As described above, all three jet algorithms are dependent on the parameter R which defines the size of the jet. In ATLAS, we usually define this R in terms of the pseudorapidity instead of rapidity:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}, \tag{4.5}$$

where η is the pseudorapidity defined in equation 2.3. In general, we distinguish between small jets with a size of $R = 0.2 - 0.4$, and large jets with sizes of $R \geq 0.8$. The small radii are used for jets originating from quarks and gluons, whereas the larger radii are used for jets formed by the hadronic decays of the W and Z bosons, the Higgs boson, and the top quark.

4.3 Small jets

Small jets are used to capture the energy deposits of individual quarks and gluons. They are constructed with the FastJet package [81] with the standard size used in ATLAS of $R = 0.4$. We distinguish between reconstructed (reco) jets, truth jets, and pile-up jets. The reco jets are defined for both data and simulated MC events and are the ones used in ATLAS analyses. They are built from topoclusters and tracks. Truth jets are formed from stable final-state particles (typically $\tau \geq 3 \times 10^{-11}$ s) in MC simulated events. This truth information represents the pure event before it interacts with the detector. The pile-up jets can be actual QCD jets originating from a pile-up vertex, or local fluctuations caused by pile-up.

4.3.1 Jet calibration

The jets need to be calibrated in order to account for pile-up, non-compensating calorimeter response, data/MC differences, and jet response dependence on several other variables. We apply the jet energy scale (JES) calibration [82] to small jets in order to restore the energy scale of reconstructed jets to the scale of simulated truth jets. The truth jets are defined as being measured at the particle-level energy scale because they are required to originate from stable, final-state particles (particles from pile-up are excluded). The JES calibration is derived from simulation and in-situ corrections based on 13 TeV data and is undertaken in stages as pictured in figure 4.4. A slightly different approach is taken for the calibration of large jets, which will be discussed in section 4.4.3.

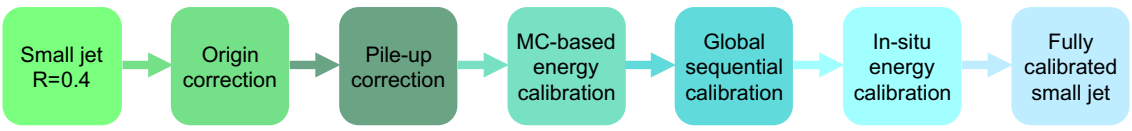


Figure 4.4: Calibration chain for small jets.

Origin correction

We start with a small jet built from topoclusters calibrated at either the EM or HAD scale as discussed in section 4.1. The first jet calibration stage is the origin correction which makes sure the jet direction is pointed to the hard scatter primary vertex instead of to the geometrical detector centre. This stage improves the η resolution of jets (measured from the difference between reco and truth jets) and does not affect the energy of the jet.

Pile-up correction

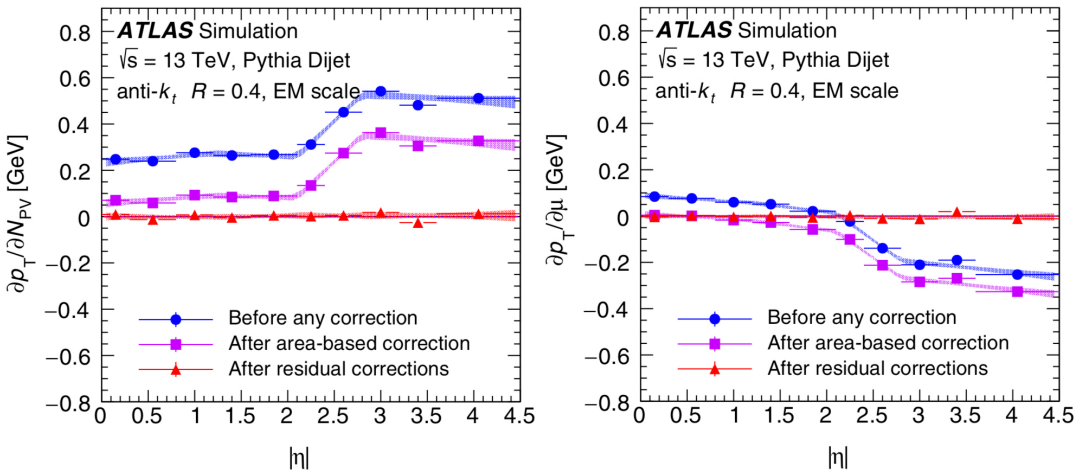
Next is the pile-up (see section 2.1.2) correction which corrects for the additional energy deposited within the jet radius due to pile-up. This additional energy is on average distributed uniformly in η and ϕ , leading to a homogeneous background that can be subtracted from individual jets [83]. The reconstructed jet p_T is corrected in two stages according to

$$p_T^{\text{corr}} = p_T^{\text{reco}} - (\rho \times A) - \alpha \times (N_{\text{PV}} - 1) - (\beta \times \mu), \tag{4.6}$$

where p_T^{reco} is the original p_T of the reconstructed jet before any corrections. The second term in equation 4.6 subtracts the pile-up fraction of the jet p_T on a per-event basis according to the jet area (A) [84]. The jet's median transverse momentum density, ρ, is used to calculate the pile-up fraction, as this is dependent on the number of reconstructed primary vertices, N_{PV}, which is sensitive to in-time pile-up. This correction is applied to the jet four-momentum and does not affect its position in η, ϕ space. The third and fourth terms in equation 4.6 represent the subtraction of the residual dependence on N_{PV} and the mean number of interactions per bunch crossing, $\langle\mu\rangle$ (sensitive to out-of-time pile-up). This dependence is calculated as the difference between the p_T of the reconstructed jet and that of the truth jet. The α and β coefficients are calculated in bins of $|\eta|$ from a logarithmic fit at $p_T^{\text{truth}} = 25$ GeV because pile-up is most relevant in this p_T region. The final results of the two pile-up correction stages are shown in figure 4.5.

MC-based energy calibration

After the pile-up correction, an MC-based energy calibration corrects the four-momentum of the jet to the particle-level energy scale. This correction is calculated from the difference between the reconstructed- and truth-jet energy (the jets are matched geometrically within $\Delta R = 0.3$). On top of this, a correction for the η direction of the jet is applied in specific detector regions where a bias in η is observed. This bias usually occurs in jets that span across different calorimeter regions with different geometry, technology, or granularity. This leads to a variety in energy responses ($E^{\text{reco}}/E^{\text{truth}}$) between calorimeter regions which can cause artefacts in the jet's reconstructed energy.



(a) In-time pile-up dependence (b) Out-of-time pile-up dependence

Figure 4.5: Dependence of jet p_T on the event pile-up as a function of $|\eta|$ for $p_T^{\text{truth}} = 25$ GeV, shown before and after the pile-up correction [82]. Figure (a) shows in-time pile-up dependence (N_{PV}) and figure (b) shows out-of-time pile-up dependence ($\langle\mu\rangle$).

Global sequential calibration

In order to remove the dependence of jet reconstruction on the jet energy distribution and particle composition, the global sequential calibration (GSC) is applied to improve the JES resolution [85]. This method reduces the sensitivity to differences in the quark-jet response versus the gluon-jet response. Jets initiated by quarks typically contain hadrons with large fractions of the jet p_T that penetrate far into the detector, whereas jets initiated by gluons usually have a wider transverse distribution of lower p_T particles that do not penetrate as deeply. The calibration corrects the jet response dependence on five observables:

1. Energy fraction of the jet in the first HAD calorimeter layer
2. Energy fraction of the jet in the last EM calorimeter layer
3. Number of tracks associated to the jet with $p_T > 1$ GeV
4. Track width: the average distance between the tracks associated to the jet and the jet axis, weighted by the track p_T
5. Number of muon track segments associated to the jet

A correction to the jet four-momentum is derived for each observable by inverting the reconstructed jet response in simulated events. The corrections are derived as a function of the transverse momentum of the truth jet and the jet's $|\eta|$ position. A normalisation factor is applied to each inversion such that the average energy at each stage remains constant. The corrections to each observable are applied sequentially to the jet four-momentum in the order presented above. Any correlations between the observables are neglected. After application of the GSC, the dependence on the above variables is reduced to below 2% [82].

In-situ energy calibration

The very last stage of the JES calibration involves residual in-situ energy corrections that account for data/MC differences in the jet response. It is applied only to data and corrects the jet p_T using other well-measured, calibrated reference objects which each focus on a different p_T region. The response of forward jets ($0.8 < |\eta| < 4.5$) is calibrated to that of well-measured central jets ($|\eta| < 0.8$) using dijet events. The response of central jets is calibrated using Z boson ($20 < p_T < 500$ GeV), photon ($36 < p_T < 950$ GeV), and multijet ($300 < p_T < 2000$ GeV) events.

Final JES uncertainty

The full JES calibration yields 80 separate systematic uncertainties propagated from the individual calibrations. 67 of these come from the final stage of in-situ calibrations and account for assumptions made in event topology, MC simulation, MC sample statistics, and uncertainties propagated from the energy scales of the electron, muon, and photon. The other 13 uncertainties are derived from the other calibration steps. The full combination of all of the JES uncertainties varies between ~ 1% for a jet of 200 GeV and ~ 4.5% for a jet of 20 GeV, as can be seen in figure 4.6.

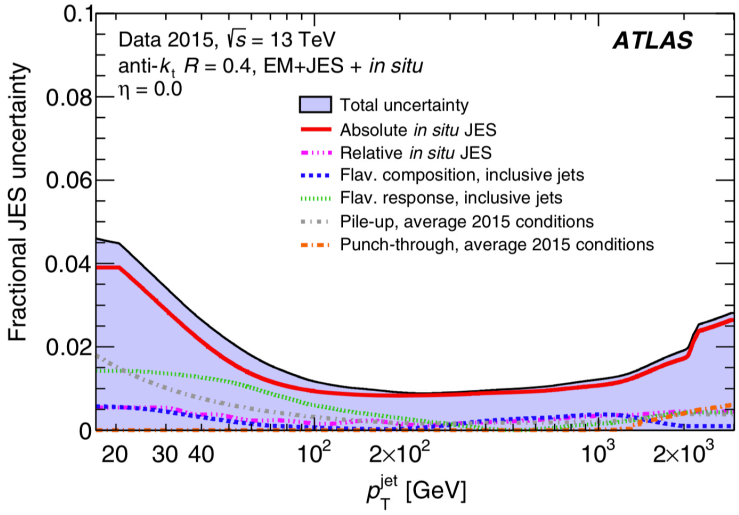


Figure 4.6: JES uncertainty as a function of the jet p_T at $\eta = 0$ [82].

4.3.2 Jet cleaning

In order to identify jets arising from non-collision sources or detector noise, certain quality criteria are imposed on the jets; this is referred to as *jet cleaning* [86]. Any event containing one or more 'bad' jets is removed from physics analyses in ATLAS. The main sources of these bad jets are beam induced background due to upstream proton losses, cosmic rays, and calorimeter noise. The jet cleaning procedure uses several jet quality variables to distinguish good jets from fake (bad) jets. These variables are based on tracking information, energy ratios, and signal pulse shapes in the LAr calorimeters. The latter type can discriminate fake jets coming from

noise in the LAr calorimeters, whereas the other types of variables are effective at rejecting fake jets from all sources. The jet cleaning proposes two selections: a *loose* and a *tight* one. The loose selection is used in most physics analyses and provides an efficiency of selecting jets from *pp* collisions above 99.5%. The tight selection provides a higher fake jet rejection with a slightly lower efficiency for good jets of 95%.

4.3.3 Rejecting pile-up jets

Pile-up jets need to be suppressed in order to correctly measure the hard scatter of interest. The rejection of pile-up jets is partially taken care of by the pile-up correction stage of the JES calibration described above, because this usually reduces the jet p_T below the p_T threshold set for jet selection in physics analyses (typically 20–25 GeV). However, some pile-up jets still remain and are rejected by use of the jet vertex tagger (JVT) discriminant [87, 88] after the JES calibration. This method is applied to both MC and data events.

The JVT algorithm uses information related to the fraction of jet p_T carried by the tracks originating from the hard scatter and is constructed based on a k-nearest neighbour algorithm. It specifically targets low-p_T jets ($p_T < 60$ GeV) in the central detector region ($|\eta| < 2.4$) which are matched to tracks with $p_T > 0.4$ GeV. QCD jets originating from a pile-up vertex are identified by the number of tracks associated to the jet. Since the ID reconstructs tracks from in-time events only, these QCD jets originating from pile-up will have little to no tracks associated to them.

The pile-up jets originating from local fluctuations, also called *stochastic jets*, are a superposition of particles from various pile-up vertices. These pile-up jets are distinguished from jets originating from the primary vertex by checking the level of stochasticity of the jets. This is measured by checking the different vertices that tracks matched to the jet come from. If the jet's tracks come from many different vertices, it is likely to be a stochastic jet and is rejected.

The default JVT cut is 0.59; if jets score below this value they are rejected as pile-up and jets scoring above this value are assumed to be from the hard scatter. This cut leads to a signal jet selection efficiency of 90% which is stable within 1% when varying N_{PV} or μ [87,88]. This means that the performance of the final JVT discriminant is independent of the amount of pile-up.

4.3.4 Flavour tagging

Many physics analyses rely on techniques that identify the original parton type of a jet in an event. This process is called *tagging* and, for small jets, one of the main tagging methods is the identification of jets containing b-hadrons, also called b-tagging. Since the b-hadrons have a relatively long lifetime ($\sim 10^{-12}$ s), they travel a significant distance of a few mm in the detector before decaying. This produces a secondary vertex with high impact parameter tracks that can be matched to jets, see figure 4.7. The impact parameter of a track is defined as the distance of closest approach to their associated primary vertex. This information is then used to distin-

guish these b-jets from jets with other flavours. As mentioned in section 2.2.2, the b-tagging performance depends critically on the operation of the ATLAS tracker. The current b-tagging algorithms rely on the new IBL detector layer installed in the 2013–2015 technical stop. The IBL optimises tracking for high pile-up and high-p_T environments.

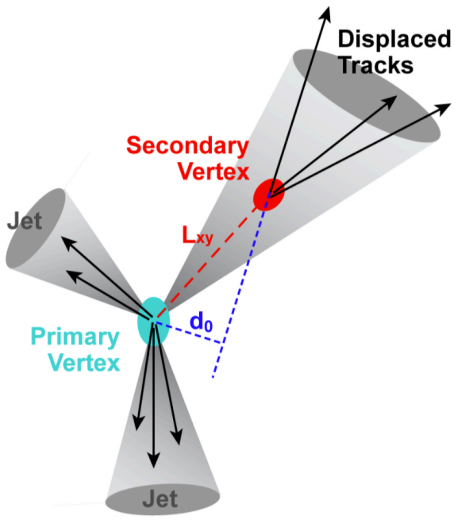


Figure 4.7: Schematic of two light jets and a b-jet displayed with their tracks. The b-jet has a large flight path, L_{xy}, which allows for the reconstruction of a secondary vertex. The displaced tracks from the b-jet have a large impact parameter, d_0.

We use a multivariate algorithm called MV2c10 [89, 90] for our b-tagging. This algorithm was trained on simulated $t\bar{t}$ events containing at least one lepton and a 7% fraction of jets containing c-hadrons (c-jets). The c-jets are a source of background for the b-jets since they leave a similar signal in the detector with a secondary vertex, even though their lifetime and thus traveled distance is shorter. The algorithm is designed to discriminate between b-jets, c-jets, jets containing hadronically decaying τ-leptons (τ-jets), and jets originating from light quarks (u, d, s) or gluons (light jets) [91].

The MV2c10 algorithm uses a BDT (see section 5.7.1 for more on BDTs) that combines 21 different variables that are sensitive to the flavour of the jet. The variable list is compiled from a combination of three different basic b-tagging algorithms based on:

- The impact parameter (IP) [91]: This algorithm uses the IP of tracks which is defined as the distance of closest approach of the track to the primary vertex. This IP is usually large for tracks coming from b-hadron decays due to their long lifetime. The IP is assigned a positive (negative) sign if the point of closest approach of the track to the primary vertex is in front of (behind) the primary vertex with respect to the jet direction. A secondary vertex behind the primary one usually indicates background events.
- The secondary vertex reconstruction [91]: This algorithm reconstructs a displaced secondary vertex inside the jet. The secondary vertex is required to have at least two tracks associated to it and is rejected if it is likely originating from a long-lived particle decay, photon conversion, or interactions with detector material. The rate of reconstructed secondary vertices is significantly higher for b-jets than c-jets or light jets.

- The decay chain multi-vertex reconstruction [92]: This algorithm aims to reconstruct the full primary vertex → b-hadron → c-hadron decay chain. It approximates the flight direction of the b-hadron and assumes that the primary vertex, the secondary vertex from the b-hadron decay, and the tertiary vertex from the c-hadron decay all lie on this line. This makes it possible to distinguish between the b-hadron decay vertex and the c-hadron decay vertex and their associated tracks.

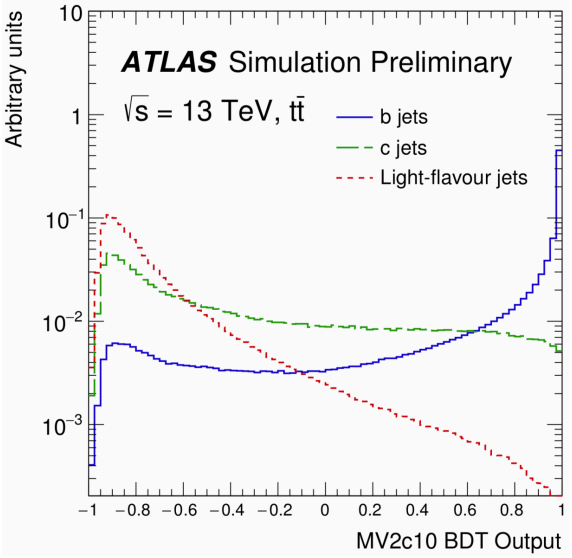


Figure 4.8: The MV2c10 BDT output score evaluated with simulated $t\bar{t}$ events for b-jets (blue), c-jets (green), and light jets (red) [90].

The final MV2c10 BDT output score is shown in figure 4.8 for b-jets, c-jets, and light jets. Higher BDT values correspond to relatively more b-jets, whereas lower output values correspond to relatively more c-jets and light jets. The algorithm is calibrated at four fixed WPs corresponding to a benchmark b-tagging efficiency and accompanying rejection factors for c, τ, and light jets, as obtained from training on simulated $t\bar{t}$ events. The b-tagging efficiency is defined as $\epsilon_b = N_b^{\text{tagged}}/N_b^{\text{true}}$ and the rejection factors are defined as the inverse of the efficiency to pass the WP, i.e. $R_c = 1/\epsilon_c$ and $R_{\text{light}} = 1/\epsilon_{\text{light}}$. The WPs are defined by a single cut value on the BDT algorithm output, see table 4.1. The BDT cut is fixed across the jet p_T spectrum, with jets required to have $p_T > 20$ GeV. The efficiency and rejection factors are optimal for the medium jet p_T range between 50–200 GeV which means that the flavour tagging does not perform as well for jets with high transverse momentum [90].

WP	BDT cut	b-jet efficiency	c-jet rejection	light jet rejection	τ-jet rejection
Very tight	0.9349	60%	34	1538	184
Tight	0.8244	70%	12	381	55
Medium	0.6459	77%	6	134	22
Loose	0.1758	85%	3.1	33	8.2

Table 4.1: Working points for the MV2c10 b-tagging algorithm and corresponding b-jet efficiency and other jet rejection rates. The values are obtained from the training on $t\bar{t}$ events with the requirement that the jet p_T is above 20 GeV. The BDT cut values are fixed across the jet p_T spectrum.

4.4 Large jets and boosted objects

Large radii are used for jets formed by the hadronic decays of the W, Z, and Higgs bosons, and the top quark. A rule of thumb for the decay of massive objects is that their decay products will be produced collimated in a cone of $R \approx 2m/p_T$, where m is the mass and p_T the transverse momentum of the particle under consideration. This means that increasing the p_T of the object decaying (also referred to as *boosting* the object) will decrease the size of the jet needed to fully capture all of its decay products. For example, the decay of a Higgs boson of $p_T = 250$ GeV can be captured in a jet of radius $R = 1.0$, whereas a jet of $R = 0.8$ suffices for a Higgs boson with a p_T of 320 GeV.

4.4.1 Jet substructure

A large jet is built of several constituents; these are generally the topoclusters discussed in section 4.1, but could also include ID tracks. These constituents can be distributed in the jet in many different ways. This inner structure of the large jet is referred to as jet substructure (JSS) and can help us to identify which particle the jet originated from. Using JSS information, we can build variables useful in jet tagging. These JSS variables are always some function of:

- The number of jet constituents
- The energy of the jet constituents
- The angular separation between the jet constituents (ΔR)

For example, looking at the location of high-p_T (hard) constituents can show whether a jet is more two-pronged (like a W boson decay) or more three-pronged (like a top quark decay). If there is no clear concentration of hard substructure, the jet is more likely to have originated from gluons and light quarks.

Jet mass

Perhaps the most intuitive and widely-used JSS variable is the jet mass. The jet mass is very useful in jet tagging since it can be compared to the mass of the object we are aiming to tag. The calorimeter jet mass is based on the jet constituent topoclusters by:

$$m^{\text{calo}} = \sqrt{\left(\sum_{i \in J} E_i\right)^2 - \left(\sum_{i \in J} \vec{p}_i\right)^2}, \tag{4.7}$$

where J is the large jet made of constituents i with energy E_i and momentum $\vec{p}_i$.

N-subjettiness

As mentioned above, it is useful to look at the pronged structure of the large jet. The JSS variable *N-subjettiness* [93, 94] precisely accomplishes this. The first step in calculating this variable is

to cluster the large jet constituents into subjets with the k_T algorithm and require that exactly N candidate subjets are found (this is also called the *exclusive* k_T algorithm). Even if the jet has less than N subjets, this algorithm forces the jet to be divided into exactly N parts. For a large jet with k constituents and N candidate subjets, the N-subjettiness variable is then defined as:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \cdot \min\left\{\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k}\right\}, \tag{4.8}$$

where $\Delta R_{j,k}$ is the distance between a candidate subjet j and constituent k. The normalisation factor d_0 is defined as $d_0 = \sum_k p_{T,k} R_0$ with R_0 the size of the large jet that we started with.

The τ_N variable tells us how much a large jet can be regarded as a jet composed of N subjets. If τ_N is close to zero it means that most of the energy in the large jet is distributed along the k_T candidate subjet directions which means it is likely to have N or fewer subjets. If τ_N is closer to one, a significant fraction of the large jet energy is misaligned with the candidate subjets and the large jet is therefore more likely to have at least $N+1$ subjets.

In general, ratios of different N-subjettiness variables are used to distinguish between large jets originating from different hadronic decays and QCD jets. For example, in order to distinguish between a W jet and jets coming from QCD, we use the ratio $\tau_{21} = \tau_2/\tau_1$. W jets will have a small τ_2 and large τ_1, but this behaviour can also be observed in QCD jets. However, those QCD jets with larger τ_1 also typically have larger values of τ_2 because these are jets composed of diffused wide angle radiation. Therefore, the τ_{21} ratio is the better discriminating variable.

k_T splitting scales

Another important substructure variable that is used frequently is the k_T splitting scale [95] (see section 4.2 for the k_T jet algorithm). If a jet was defined with one of the other jet algorithms (anti-k_T or C/A), we first recluster the jet constituents with the k_T algorithm. We use equation 4.2, with $p = 1$ to specify the k_T algorithm, and take its square root which leads to the definition of the k_T splitting scale:

$$\sqrt{d_{ij}} = \min\left(p_{T,i}, p_{T,j}\right) \times \frac{\Delta R_{i,j}}{R}. \tag{4.9}$$

To specify the first splitting scale, $\sqrt{d_{12}}$, the i, j are taken to be the two proto-jets combined at the final step of the k_T algorithm. These two proto-jets are the most widely separated and highest p_T jet constituents. Similarly, the second k_T splitting scale, $\sqrt{d_{23}}$, uses the two proto-jets from the penultimate step of the k_T clustering sequence. For a two-body heavy particle decay, the expected value of $\sqrt{d_{12}}$ is about half the mass of the particle since the final step combines the two hardest proto-jets. Jets originating from gluons or light quarks have a sharp peak at small $\sqrt{d_{ij}}$ values with a quick drop-off. In this way, the k_T splitting scales can be used to distinguish heavy-particle decays from QCD splittings.

4.4.2 Boosted object tagging

The process of identifying from which particle a jet originated is referred to as jet *tagging*. A form of jet tagging on small jets was described in section 4.3.4, but tagging methods are more

commonly applied to large jets in order to identify the boosted objects that formed them. The main focus of boosted object tagging in ATLAS is distinguishing jets coming from boosted top quarks, W bosons, and Higgs bosons from jets coming from quarks and gluons. In the boosted $t\bar{t}H(H \to b\bar{b})$ analysis, a form of top-tagging is studied which is based on the JSS variables jet mass and N-subjettiness, as described above. The N-subjettiness ratio $\tau_{32} = \tau_3/\tau_2$ is used because it can distinguish large jets with a three-pronged structure, which is what we would expect from a hadronically decaying top quark, from jets with a two-pronged structure, which we expect from a W boson decay.

The top tagger [96] is tested for signal in $Z' \to t\bar{t}$ events and for background in multijet events. The algorithm provides two WPs at 50% and 80% signal efficiency, which are the *tight* and *loose* WPs, respectively. The cuts on the mass and the τ_{32} variable are optimised per large jet p_T bin. For the 50% WP, the upper allowed value of τ_{32} ranges from 0.75 at a jet p_T of 200 GeV to 0.57 for $p_T \geq 1600$ GeV. At the 80% WP, this range is 0.85–0.7 for the same p_T cuts. As the jet p_T increases, the decay products of the top get more collimated and are therefore better contained inside the large jet. We thus expect a clearer three-pronged structure with increasing p_T which leads to lower values of τ_{32}. The lower mass threshold for the 50% WP varies from 85 GeV at a large jet p_T of 200 GeV, to 140 GeV for $p_T \geq 1600$ GeV. For the 80% WP these thresholds are decreased to 70 GeV and 135 GeV respectively. Again, this is because the tops are not fully contained at the lower end of this p_T spectrum and we will therefore not catch their full mass in one large jet.

4.4.3 Large jet calibration and grooming

The topoclusters used to construct large jets are calibrated at the hadronic scale using the LCW scheme (see section 4.1). The large jets need to undergo a calibration procedure for the same reasons as the small jets. The calibration chain for large jets is shown in figure 4.9.

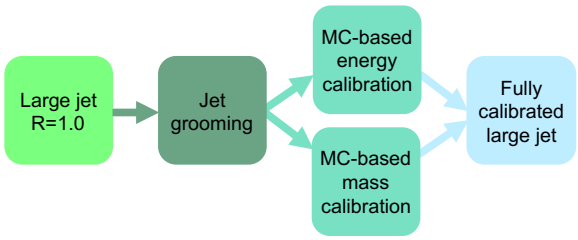


Figure 4.9: Calibration chain for large jets.

Jet grooming

The first step in calibrating large jets is *jet grooming* which gets rid of soft and wide-angle radiation. The grooming of jets reveals their pronged substructure and thereby improves the resolution of JSS variables which are used in jet tagging. The jet grooming also decreases the effects of pile-up and makes the jet calibration easier and more accurate. This is the reason why large jets have a simpler calibration procedure than small jets.

There are various ways to groom a jet; the standard used in ATLAS is jet *trimming* [97]. The trimming procedure starts with the reclustering of the large jet constituents into subjets of size R_{sub} with the k_T algorithm. In order to get rid of soft radiation, we define a subjet p_T threshold relative to the total jet p_T (f_{cut}). We remove all subjets with $p_{T,i}/p_T^{\text{jet}} < f_{\text{cut}}$ and the constituents making up the remaining subjets form the trimmed large jet. In ATLAS, the standard parameters chosen are $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$ which was determined to be the optimal configuration in reference [98]. An illustration of the trimming procedure is shown in figure 4.10.

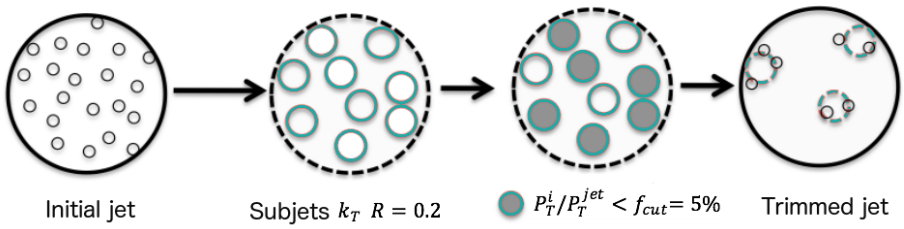


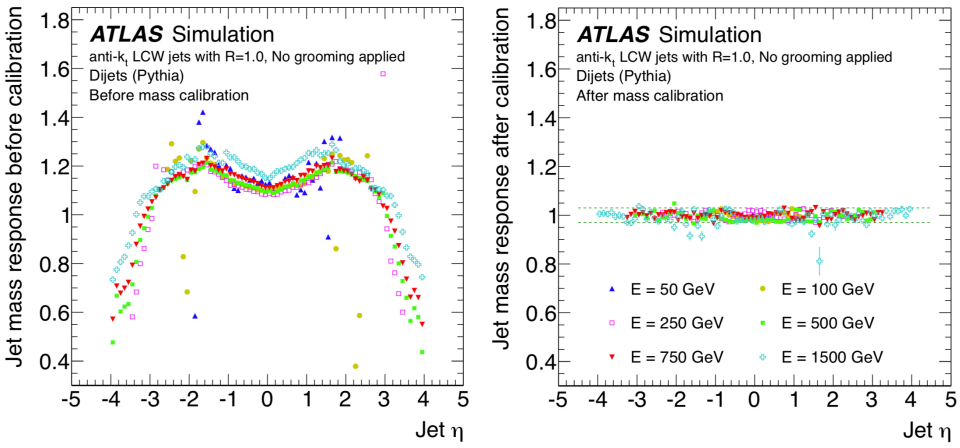
Figure 4.10: Illustration of the jet grooming technique known as trimming.

The trimming procedure can remove small parts of radiation from the hard scatter and final-state radiation (FSR), but typically removes soft contributions from pile-up, UE, multi-parton interactions (MPI), and initial-state radiation (ISR). The removed mass fraction is much larger for jets originating from light quarks or gluons than for jets originating from boosted particles. In this way, trimming makes it easier to distinguish between these types of jets.

MC-based calibration

The final step in the large jet calibration chain is based on MC studies. This step involves both a jet energy scale (JES) calibration and a jet mass scale (JMS) calibration. The energy calibration follows the same procedure as was described for the small jet calibration in section 4.3.1: the four-momentum of the jet is corrected to the particle-level energy scale and the η direction of the jet is corrected in detector regions where a bias in η is observed. The energy response is defined as $R_E = \langle E_{\text{reco}}/E_{\text{truth}} \rangle$, where E_{reco} is the energy of the uncalibrated reco jet, and E_{truth} is the energy of the corresponding particle-level jet. The brackets denote that the response is specified by the mean value of the response distribution. A correction factor is extracted from this mean [99].

After the JES correction, the JMS correction is applied in the same way, using its own response distribution $R_m = \langle m_{\text{reco}}/m_{\text{truth}} \rangle$. The jet mass response is shown as a function of jet η in figure 4.11 before (a) and after (b) the JMS calibration is applied (the JES calibration is already applied in both cases). It is evident that the jet mass is very sensitive to soft, wide-angle radiation and hence the mass calibration is very important. Without this calibration, the mass of large jets in the central detector region can differ up to 20% from the particle-level true jet mass, whereas this difference is even larger for the non-central jets.



(a) Before JMS calibration (b) After JMS calibration

Figure 4.11: Jet mass response $R_m = \langle m_{\text{reco}}/m_{\text{truth}} \rangle$ for ungroomed anti-k_T jets with $R = 1.0$, shown before (a) and after (b) the JMS calibration is applied [99].

4.4.4 Reclustered jets

In the boosted $t\bar{t}H$ analysis, discussed in chapter 5, we have traditionally used large jets built directly from topoclusters calibrated at the hadronic scale, using the anti-k_T algorithm and a jet radius of $R = 1.0$. In order to remove soft and wide-angle radiation from these large jets, the trimming procedure is applied with $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$. The disadvantage of using these trimmed large jets (also called standard large jets from now on) is that they bring a large extra systematic uncertainty to the analysis. These uncertainties are large because they are calculated with the double r-track ratio method, as detailed in reference [100]. I have introduced another method of defining large jets in the boosted $t\bar{t}H$ analysis: the reclustering method [101]. In this method, the topoclusters in each event are first clustered into small $R = 0.4$ anti-k_T jets and calibrated with the full JES calibration as described in section 4.3.1. Afterwards, these small jets are used as inputs to any sequential recombination scheme (such as anti-k_T, k_T, or Cambridge/Aachen) in order to construct a large jet. An illustration of the reclustering method is shown in figure 4.12.

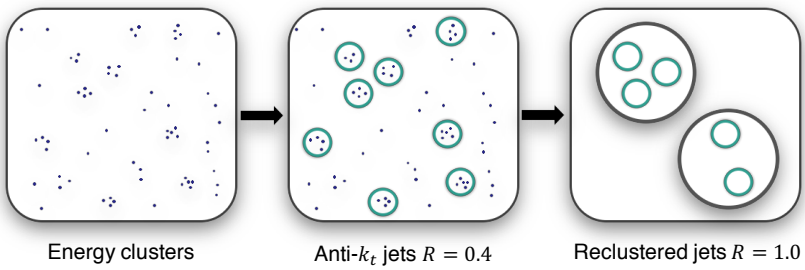


Figure 4.12: Illustration of building large jets from smaller jets with the reclustering technique.

The anti-k_T $R = 0.4$ jets that a reclustered jet is built from are called its *subjets*. They are fully calibrated and the corrections, calibrations, and uncertainties directly propagate from the small to the large jet. This leads to smaller systematic uncertainties than the trimmed jets and

does not require calculating the specific large jet uncertainties and calibrations. A recent study of in-situ measurements [102] confirms that the data/MC differences observed with reclustered jets are indeed covered by simply propagating the uncertainties associated with their anti-k_T subjets. One other advantage of using reclustered jets is that we are free to choose any value for the large jet radius, any jet clustering algorithm, and different jet grooming strategies, because we do not need an additional calibration besides the small jet JES calibration already applied. For standard jets we can only use those parameters that have a full large jet calibration available, which usually means using trimmed anti-k_T $R = 1.0$ jets.

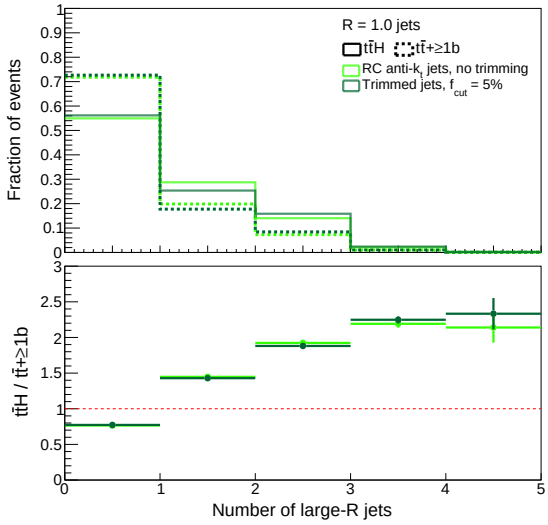
4.5 Reclustered jets studies for boosted $t\bar{t}H$ analysis

Several studies are carried out in order to understand the differences between the standard trimmed jets and reclustered jets, and to assess the best reclustered jet configuration for the $t\bar{t}H$ analysis. To this end, the large jets are studied in simulated events of both $t\bar{t}H$ and $t\bar{t}+\geq 1b$ samples, since $t\bar{t}+\geq 1b$ is the most difficult background to distinguish from the signal (see section 5.4.2). In this section, we compare the performance of reclustered jets with that of standard large jets and test the effects of applying the trimming technique on the reclustered jets. Studies are performed in order to choose the optimal jet algorithm and jet radius for the boosted $t\bar{t}H$ analysis.

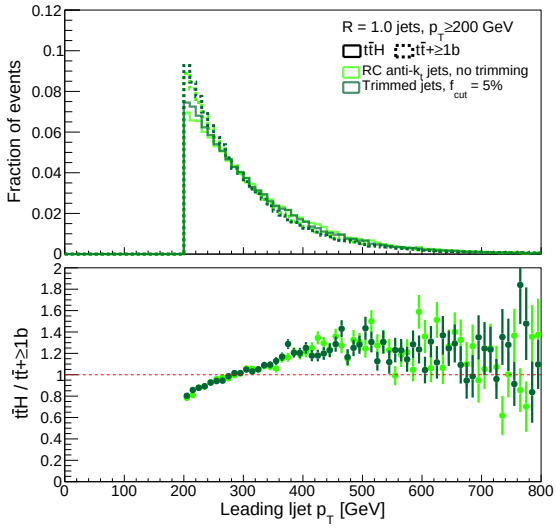
A loose event selection is applied in which we require events to have at least one lepton (electron or muon) with $p_T > 20$ GeV, and at least four small anti-k_T $R = 0.4$ jets of which at least two are b-tagged at the tight (70%) working point. All small jets are required to have a transverse momentum of at least 25 GeV after they are calibrated to the truth JES. This selection is the standard baseline selection used to get a sample of mostly $t\bar{t}$ and $t\bar{t}H(H \to b\bar{b})$ events. It is chosen loose enough in order to have significant statistics for this study. No additional cuts for a boosted selection are chosen here in order to have an unbiased study of a broad range of events. All large jets are required to have a mass > 50 GeV, a transverse momentum between 200 and 1500 GeV, and $|\eta| < 2$. The reclustered jets are required to have at least two subjets and the standard large jets are trimmed with $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$.

4.5.1 Reclustered vs. trimmed jets

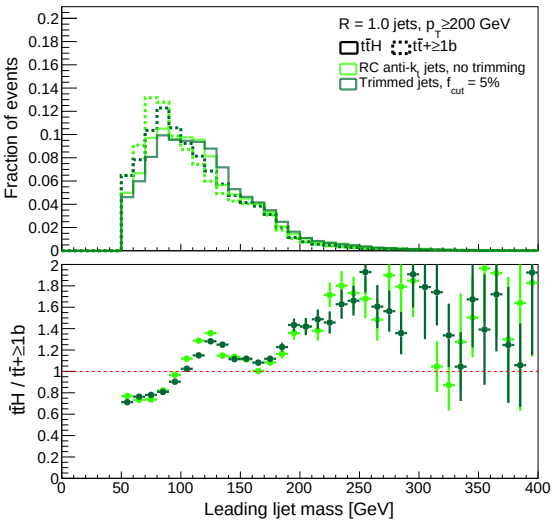
In order to use reclustered jets instead of trimmed jets in the boosted $t\bar{t}H$ analysis, studies are performed to check the differences in performance and kinematics between these two types of large jets. The results of the study comparing reclustered jets to trimmed jets are summarised in figures 4.13 and 4.14. The ratio plots compare the $t\bar{t}H$ and $t\bar{t}+\geq 1b$ samples in order to assess whether a good separation between signal and background can be accomplished. All the plots consider the highest-p_T large jet, also called the *leading* jet. In order to have a consistent comparison, both the reclustered and trimmed jets use the anti-k_T algorithm and have a jet radius of $R = 1.0$. We can see that the large jet multiplicity 4.13(a), p_T 4.13(b), and mass 4.13(c) are very similar between the two types of jets. The distributions are almost equal in shape,



(a)

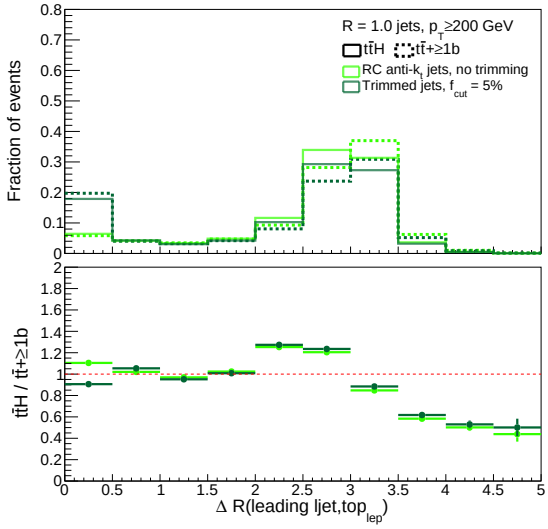


(b)

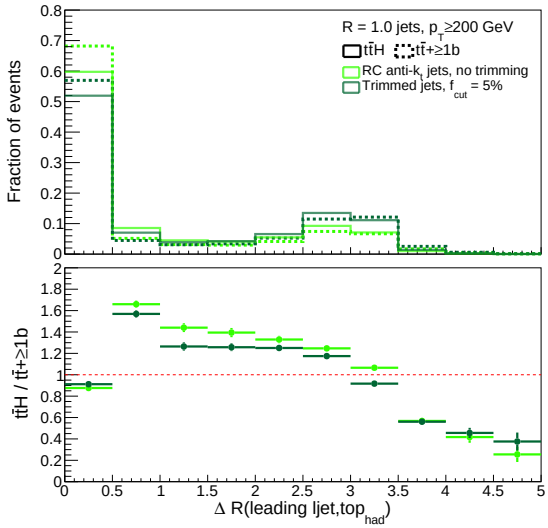


(c)

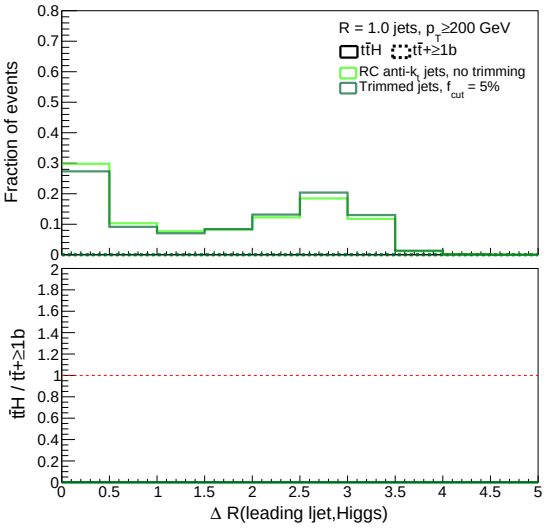
Figure 4.13: Distributions of $R = 1.0$ large jet variables comparing standard trimmed jets with $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$ and reclustered jets using the anti-k_T algorithm. The $t\bar{t}H$ signal events are shown in solid lines and the $t\bar{t}+ \geq 1b$ background events in dashed lines.



(a)



(b)



(c)

Figure 4.14: Distributions of $R = 1.0$ large jet variables comparing standard trimmed jets with $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$ and reclustered jets using the anti-k_T algorithm. The $t\bar{t}H$ signal events are shown in solid lines and the $t\bar{t}+\geq 1b$ background events in dashed lines. The ΔR distances are measured with respect to the truth level hadronic top quark, leptonic top quark, and Higgs boson.

as are the separations between $t\bar{t}H$ and $t\bar{t}+\geq 1b$ events. We distinguish a clear peak in $t\bar{t}H$ events compared to $t\bar{t}+\geq 1b$ events in the large jet mass around the Higgs mass of ~ 125 GeV.

We also look at the matching of the large jets to truth particles in the event. Specifically, we study the leptonic top, hadronic top, and the Higgs boson. The jet matching is checked using the distance ΔR between the leading jet and the particle at MC truth level. In figure 4.14(a), we see that the distance between the leading jet and the leptonic top is consistent between the trimmed- and reclustered-jets. In 4.14(b), the distance between the leading jet and the hadronic top is shown, where it is striking that in more than half of the events the leading jet is matched to the truth hadronic top within $\Delta R < 0.5$. The behaviour for trimmed and reclustered jets is quite similar, though the reclustered jets show a larger distinction between $t\bar{t}H$ and $t\bar{t}+\geq 1b$ jets in the region $0.5 < \Delta R < 3$. The ratio plot for the ΔR(leading ljet, Higgs) shown in 4.14(c) is not filled since there are no truth Higgs bosons in the $t\bar{t}+\geq 1b$ sample. Again, the behaviour is similar for trimmed and reclustered jets. From these studies we conclude that it is safe to use reclustered jets in the $t\bar{t}H$ analysis and we do not expect significant differences w.r.t. trimmed jets.

4.5.2 Trimming applied on reclustered jets

The effect of applying the trimming technique to the reclustered jets is studied; the results are shown in figure 4.15. Four different levels of trimming are compared with $f_{\text{cut}} = 5\%, 10\%, 15\%, 20\%$, and 0% to show the baseline without trimming. Instead of clustering the subjet topocluster constituents into $R = 0.2$ k_T jets (as done in section 4.4.3), we apply the trimming cut directly to the anti-k_T $R = 0.4$ subjets themselves; therefore $R_{\text{sub}} = 0.4$. In figure 4.15 we compare the results of these different trimmings cuts for (a) the large jet multiplicity, (b) the subjet multiplicity, (c) the transverse momentum, and (d) the mass of the leading reclustered jet. We do not see any significant effect from applying trimming to the reclustered jets. This is due to the fact that the reclustered jets have an inherent effective grooming since the small jets are required to have $p_T > 25$ GeV after application of the JES calibration. It is therefore decided not to use any trimming on our reclustered jets in the $t\bar{t}H$ analysis.

4.5.3 Choosing a jet algorithm

Since we are free to choose the jet algorithm used to recluster our large jets from our small jets, a comparison between the anti-k_T, k_T, and C/A algorithms is made. The results are shown in figure 4.16 for (a) the large jet multiplicity, (b) the subjet multiplicity, (c) the transverse momentum, and (d) the mass of the leading reclustered jet. We maintain the same distributions as seen in the previous figures, and observe no significant differences between the three sequential recombination algorithms. For this reason, we choose to stick with the anti-k_T algorithm for our large reclustered jets.

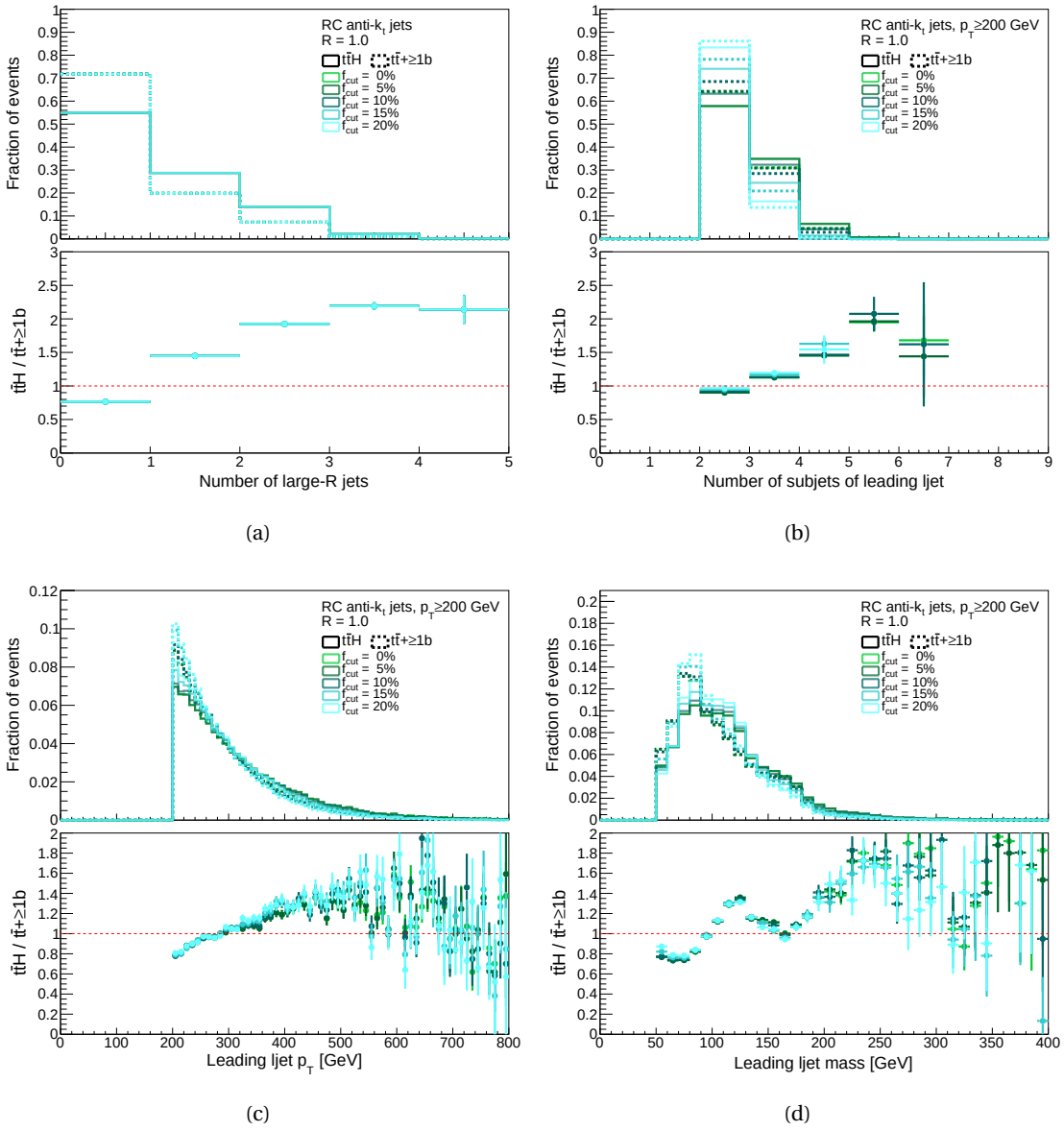


Figure 4.15: Distributions of reclustered $R = 1.0$ anti-k_T jet variables with varying levels of trimming applied. The trimming is directly applied to the anti-k_T $R = 0.4$ subjets of the reclustered jets, therefore $R_{\text{sub}} = 0.4$.

4.5.4 Choosing a jet radius

As mentioned above, the reclustering method allows us to choose any large jet radius since we do not need an extra large jet calibration. Since the decay products of a boosted object are produced in a distance of $R \approx 2m/p_T$ from each other, we expect that a larger jet radius can capture objects that are less boosted, whereas highly boosted objects can be captured fully in a smaller radius jet. In order to choose a suitable jet radius for the boosted $t\bar{t}H$ analysis, we study six different values of R for reclustered anti-k_T jets: 0.8, 1.0, 1.2, 1.4, 1.6, and 1.8.

Figure 4.17 shows comparisons for the six different jet radii of the large jet multiplicity and the leading jet's subjet multiplicity. The large jet multiplicity is shown in subfigure (a) and shows that, for all jet radii, the $t\bar{t}H$ sample more frequently has two large jets than the

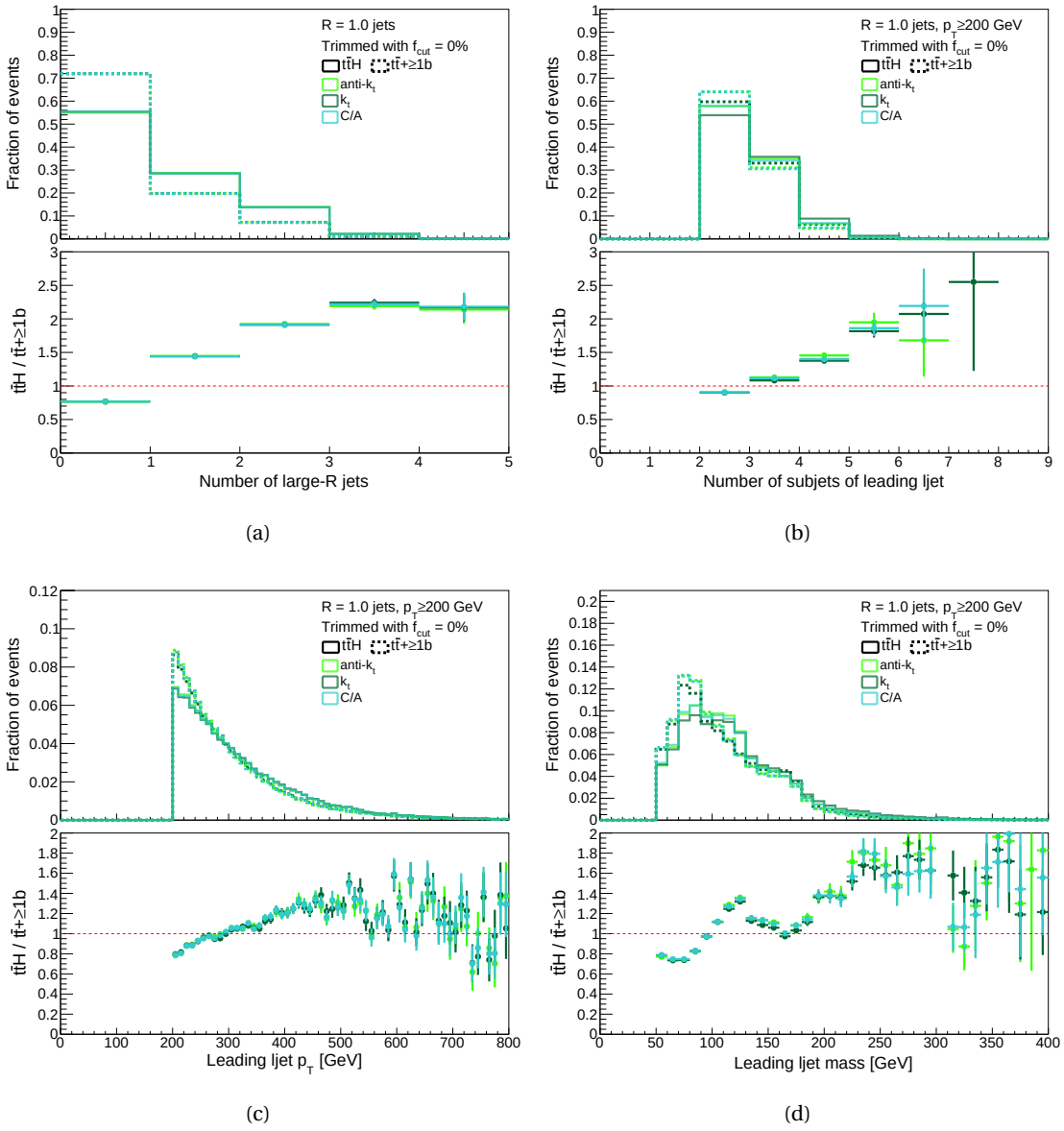
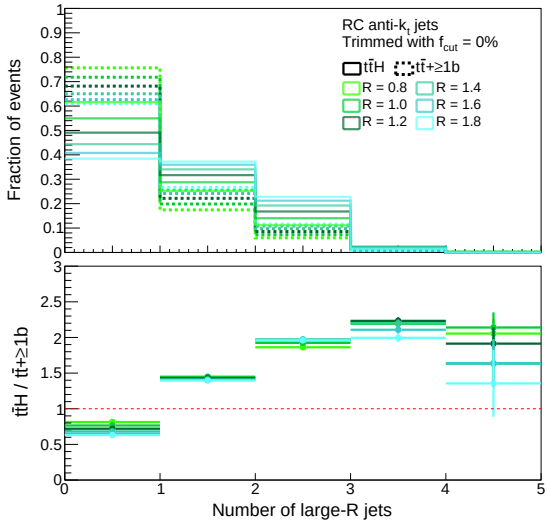


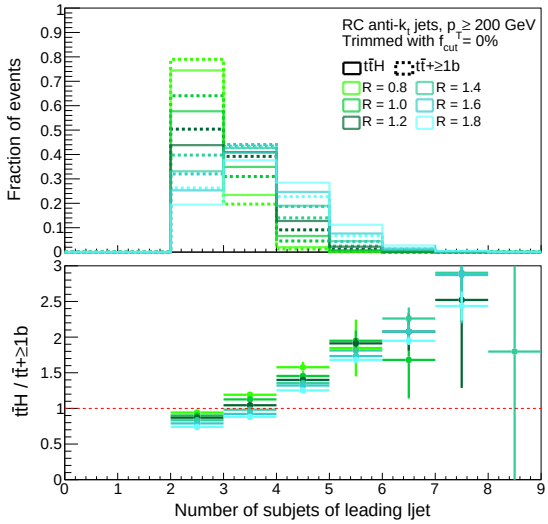
Figure 4.16: Distributions of reclustered $R = 1.0$ jet variables comparing the three different jet algorithms.

$t\bar{t}+ \geq 1b$ sample. This distinction can be beneficial in the design of a signal region for the boosted $t\bar{t}H$ analysis since requiring at least two large jets can already cut away a part of the difficult $t\bar{t}+ \geq 1b$ background. In subfigure (b) we see that the number of subjets increases along with an increase in jet radius, which is as expected. The distributions of the b-tagged subjet multiplicity (c) look very similar for all jet radii.

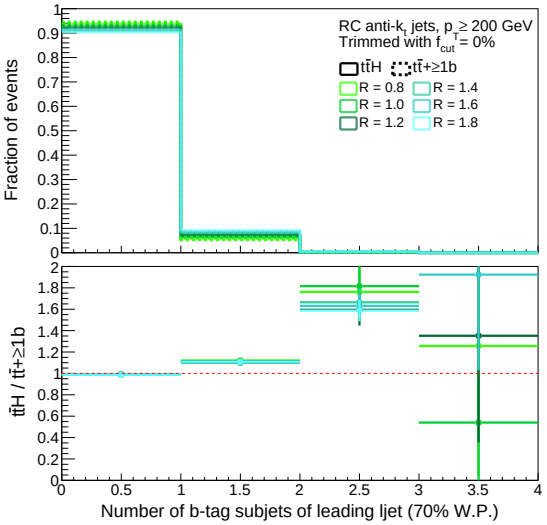
The leading jet kinematics are shown in figure 4.18. The transverse momentum of the jet in subfigure (a) looks very similar for all jet radii. The leading large jet mass shown in (b) shows clearly the shift towards higher masses when increasing the jet radius. This is expected since a larger jet has a larger catchment area and therefore more subjets, as illustrated in figure 4.17(b), which contribute to a higher mass. We see a peak around the Higgs mass ($\sim$ 125 GeV) and a dip around the top quark mass ($\sim$ 173 GeV) in the ratio plot. These features are more pronounced for the smaller radius jets than the larger ones. This is likely due to the fact that a too large jet



(a)

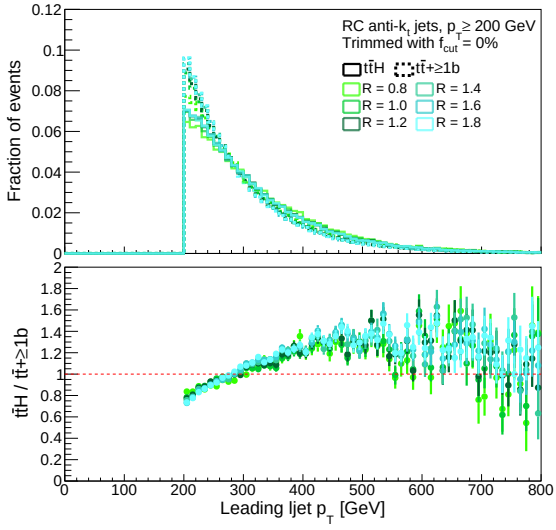


(b)

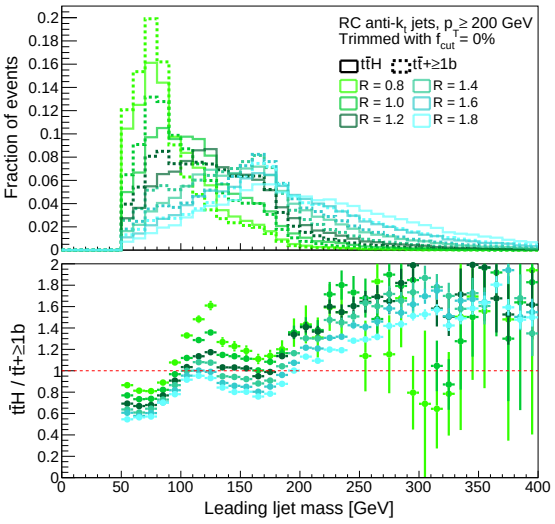


(c)

Figure 4.17: Comparisons of six different jet radii for the distributions of the reclustered anti-k_T jet multiplicity (a), the leading reclustered jet subjet multiplicity (b), and the leading reclustered jet b-tagged subjet multiplicity (c).



(a)

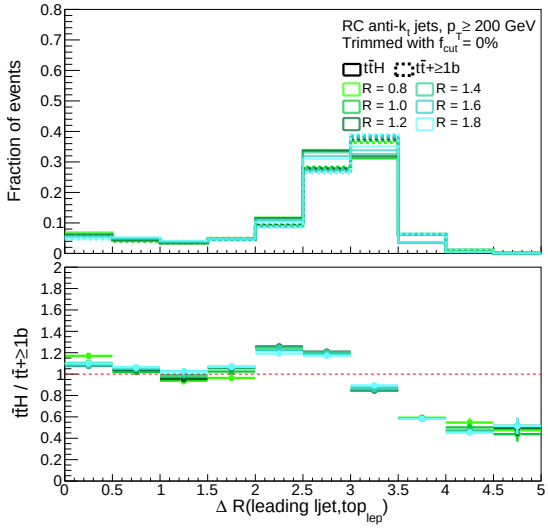


(b)

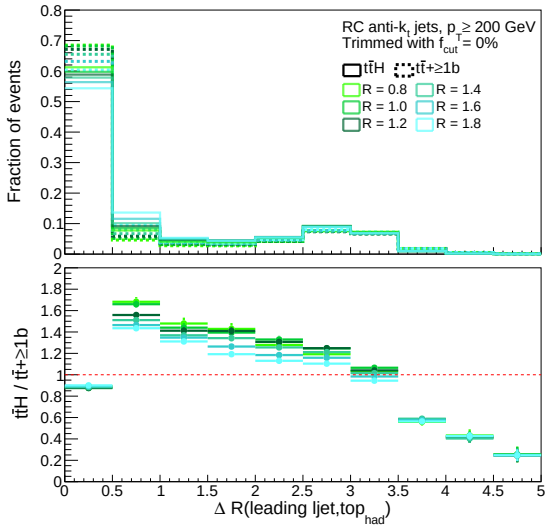
Figure 4.18: Comparisons of six different jet radii for the distributions of the leading reclustered jet transverse momentum (a) and jet mass (b).

radius will capture more soft and wide-angle radiation which degrades the mass peak resolution.

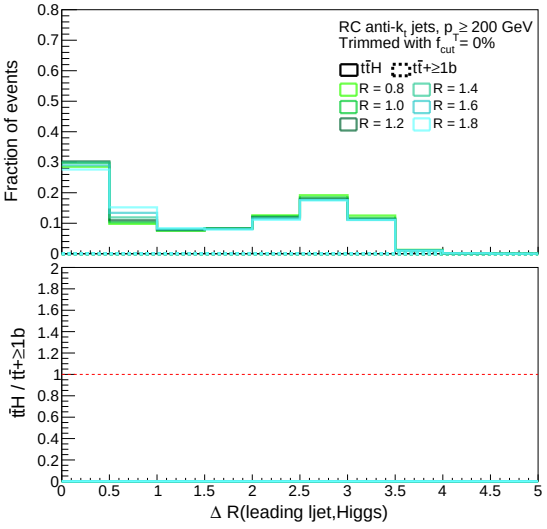
Figure 4.19 shows the distance between the leading large jet and MC truth particles when varying the jet radius. In figure (a), we see that the leading jet is often produced far away from the truth leptonic top. It is produced back-to-back ($\Delta R = \pi$) with the leptonic top jet slightly less often in the $t\bar{t}H$ sample than in the $t\bar{t}+\geq 1b$ sample. It is expected that the leptonic top is not the leading jet since there is always some MET from the neutrino in its decay. Subfigure (b) shows that in $60-70\%$ of the events the leading jet is matched to the truth hadronic top within $\Delta R < 0.5$ and thus that the hadronic top is usually the hardest object in the event. The leading jet is matched to the truth Higgs boson in about 30% of $t\bar{t}H$ events, as shown in (c). The ratio plot in (c) is not filled since there are no truth Higgs bosons in the $t\bar{t}+\geq 1b$ sample. Overall, we see very similar behaviour for all jet radii for these distance measurements.



(a)



(b)



(c)

Figure 4.19: Distributions of the distance between the leading reclustered anti-k_T jet and the truth leptonic top quark (a), the truth hadronic top quark (b), and the truth Higgs boson (c). A comparison is made between six different jet radii.

In order to check what kind of object we are catching in our jets, the subjet multiplicity is plotted against the large jet mass for $t\bar{t}H$ events in figure 4.20 and for $t\bar{t}+\geq 1b$ events in figure 4.21. We expect the Higgs jets to have two subjets (one for each of the b-quarks in its decay) and a mass close to 125 GeV. The hadronic top is expected to have three subjets (one b-quark from the first stage in its decay and two other quarks from the decay of the W boson) and a mass close to 173 GeV. The figures compare anti-k_T reclustered jets with a p_T cut of 200, 250, and 300 GeV and a jet radius of R=1.0, 1.2, and 1.4. These values for R are chosen from figure 4.18(b) which shows that the difference between the $t\bar{t}H$ and $t\bar{t}+\geq 1b$ samples in the leading jet mass is larger for smaller jet radii which is beneficial for our signal-to-background separation. A radius of 0.8 is excluded because this shows a clear jet mass peak around 80 GeV which means that we are picking up only part of the top decay corresponding to the W boson.

For both the $t\bar{t}H$ and $t\bar{t}+\geq 1b$ samples we see the trend that larger jet radii lead to a larger jet mass and a higher multiplicity of subjets. Increasing the p_T cut on the leading jet removes some events from the sample which leads to an overall lower density in the plots on the second and third rows. However, we can see that the increase in p_T leads to a higher fraction of jets with a larger mass and more subjets. In both samples, the leading large jet looks more often like a Higgs jet for jets with $R = 1.0$ or $R = 1.2$ and a transverse momentum above 200 or 250 GeV. The leading jet looks more often like a top jet for jets with $R = 1.2$ or $R = 1.4$ and $p_T > 250$ or $p_T > 300$ GeV.

Since we conclude from figure 4.18(b) that the difference between the $t\bar{t}H$ and $t\bar{t}+\geq 1b$ samples in the leading jet mass is larger for smaller jet radii, we would like to pick a jet radius that is not too large. Together with the information from figures 4.20 and 4.21, a large jet radius of 1.0 is chosen as our standard for the boosted $t\bar{t}H$ analysis. This radius allows us to capture both Higgs and top jets in our signal sample, as concluded from subfigures 4.20 (a), (d), and (g).

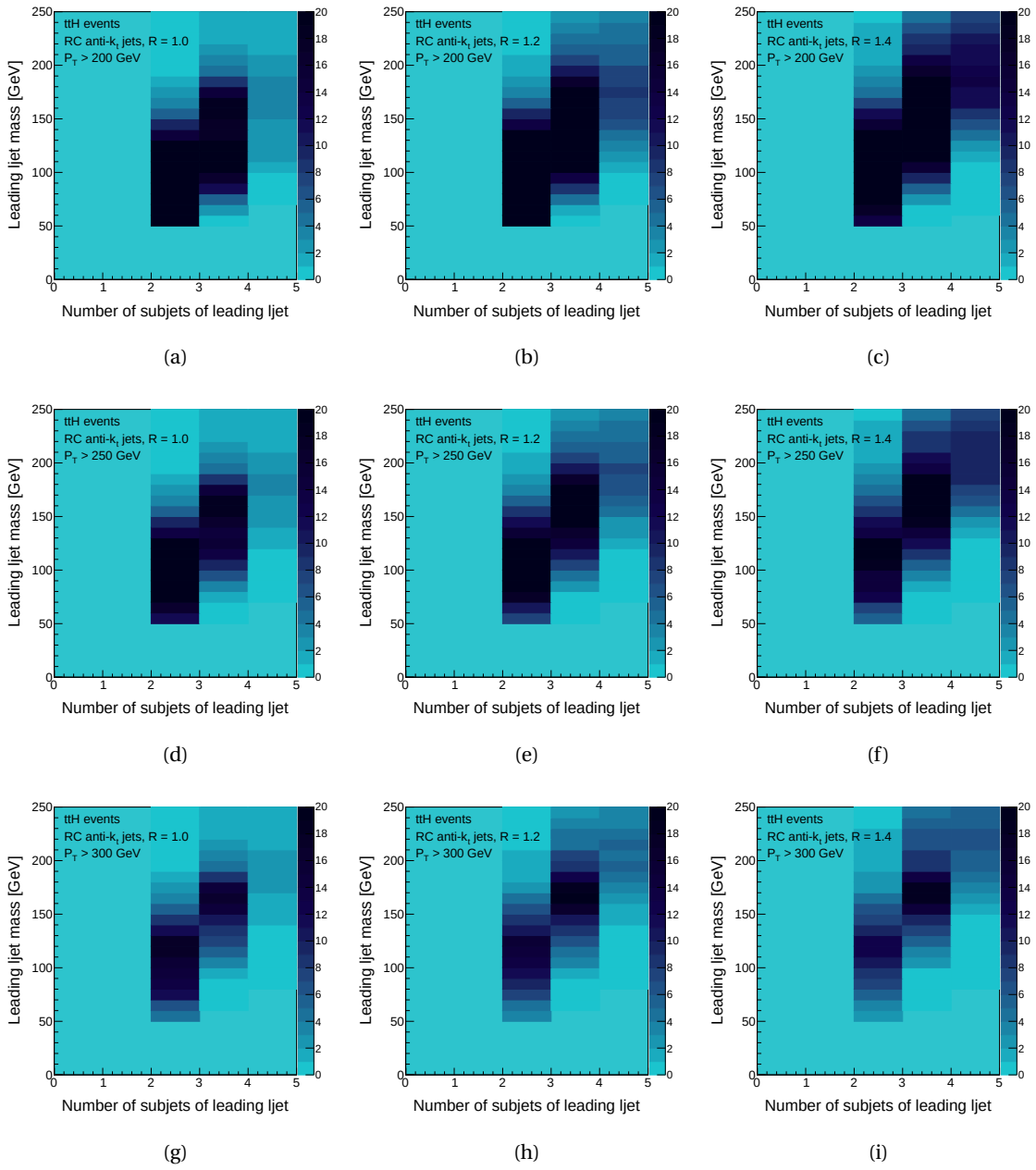


Figure 4.20: Distributions of the subjet multiplicity plotted against the large jet mass for $t\bar{t}H$ events, using anti-k_T reclustered jets. No trimming is applied on the jets. The first row has a p_T cut applied to the large jet of 200 GeV, the second row of 250 GeV, and the third of 300 GeV. The first column shows reclustered jets with a radius of $R = 1.0$, the second of $R = 1.2$ and the third of $R = 1.4$.

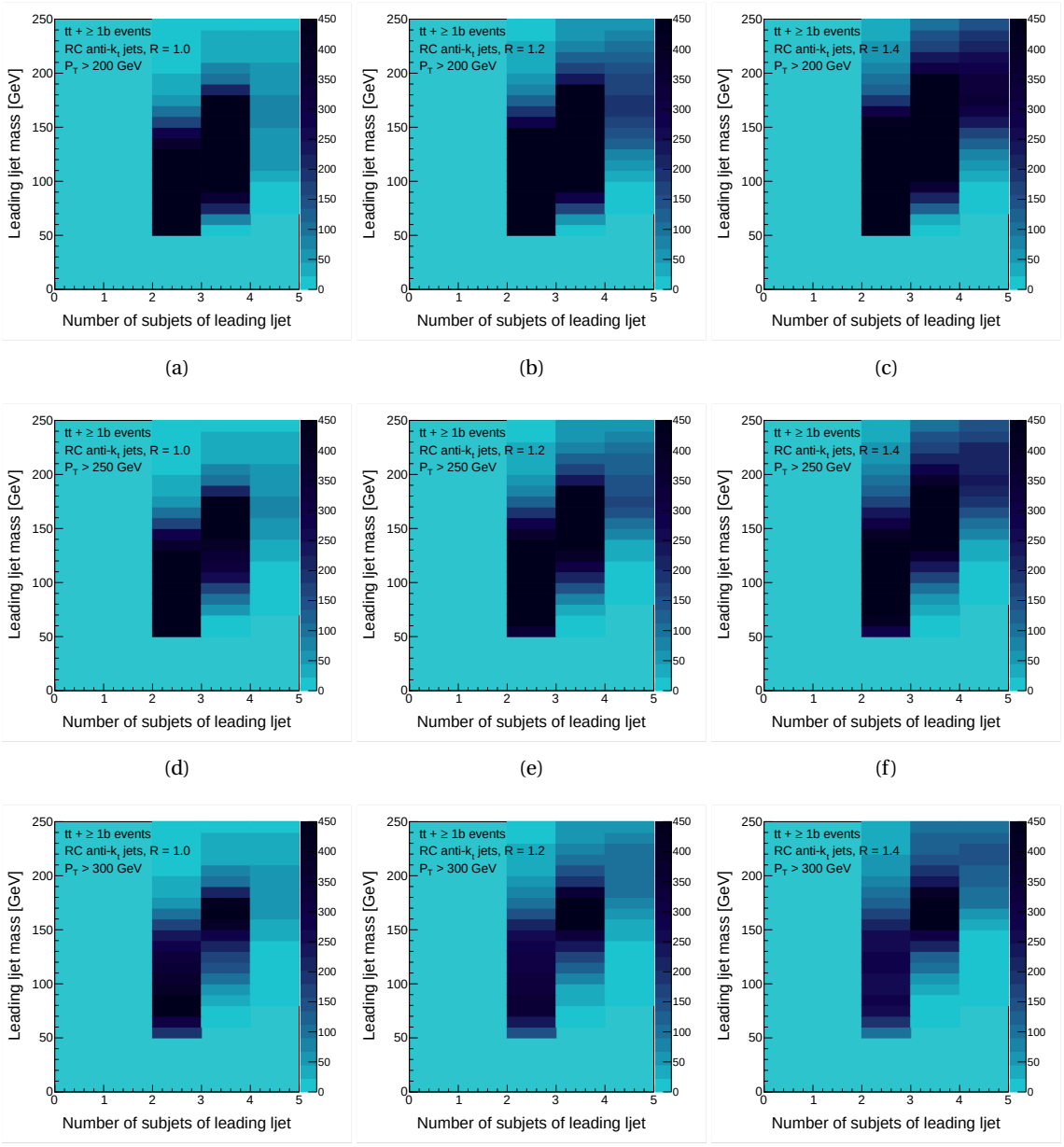


Figure 4.21: Distributions of the subjet multiplicity plotted against the large jet mass for $t\bar{t}+\geq 1b$ events, using anti-k_T reclustered jets. No trimming is applied on the jets. The first row has a p_T cut applied to the large jet of 200 GeV, the second row of 250 GeV, and the third of 300 GeV. The first column shows reclustered jets with a radius of $R = 1.0$, the second of $R = 1.2$ and the third of $R = 1.4$.

5 THE $t\bar{t}H(H \to b\bar{b})$ ANALYSIS STRATEGY

This chapter describes the analysis searching for $t\bar{t}H(H \to b\bar{b})$ in ATLAS. The techniques and strategies here are published in reference [103] and are based on the full 2015–2016 dataset from ATLAS at $\sqrt{s} = 13$ TeV which amounts to an integrated luminosity of 36.1 fb^{-1}. This search builds on previous searches for the same process performed with ATLAS data recorded at $\sqrt{s} = 7$ TeV [24] and 8 TeV [25, 26].

The full analysis strategy is described with focus given to the boosted single-lepton channel. The event selection in the analysis is based on the number of (large) jets and the number of b-tagged jets at various working points. The selection is designed to select $t\bar{t}H$ events where the Higgs decays to two bottom quarks, but all Higgs decay modes selected are treated as signal. The analysis uses multivariate analysis (MVA) techniques in order to distinguish signal events from background events.

5.1 Motivation

All current Higgs measurements performed at the LHC have been consistent with the SM [104]. One interesting place to look for new physics is in the top quark Yukawa coupling strength. As described in section 1.2, the Yukawa couplings are proportional to the mass of the fermion. Since the top quark is the heaviest particle in the SM, we expect its coupling to the Higgs to be the largest of all the fermions; it is therefore an ideal candidate to show signs of new physics. The $t\bar{t}H$ process gives a unique direct probe of this Yukawa coupling. The search is designed for the $H \to b\bar{b}$ decay channel since it has the largest branching ratio (BR) in the SM of ~ 58% for a Higgs mass of 125 GeV.

At the current centre-of-mass energy of 13 TeV, Higgs bosons can be produced at transverse momenta well above their rest mass, which means that they can be probed in the boosted regime. This region of phase space will become more and more important as the energy of collisions increases further in Run III of the LHC and at the High Luminosity LHC (HL-LHC). It is therefore important to study the techniques and sensitivities of the boosted $t\bar{t}H$ channel at this stage, in order to be ready for the future dataset with a larger contribution of boosted $t\bar{t}H$ events. Once the channel acquires enough statistics, it can be used for a differential

cross-section measurement in the high-p_T Higgs phase space. The boosted $t\bar{t}H$ channel also provides a good testing ground for studying interesting and innovative boosted techniques for both the Higgs boson and the top quark.

Besides it being an interesting channel to study for the above reasons, the boosted channel can potentially improve the sensitivity of the inclusive $t\bar{t}H(H \to b\bar{b})$ analysis. The final state of this process is composed of many jets stemming from the Higgs boson and top quark decay products, as well as from additional radiation. Many combinations of these jets are possible when reconstructing the Higgs boson and top quark candidates. The boosted regime combines several of the final state jets into large jets and therefore has the advantage of a simplified combinatorial background compared to the resolved regime. This can help to increase the purity of signal regions and thereby improve the sensitivity of the analysis.

5.2 Analysis overview

The $t\bar{t}H(H \to b\bar{b})$ search in ATLAS is split into two channels: the semileptonic channel in which one of the tops in the $t\bar{t}$ system decays leptonically and the other hadronically, and the dilepton channel in which both of the tops decay leptonically. Examples of the tree-level Feynman diagrams for these processes are shown in figure 5.1.

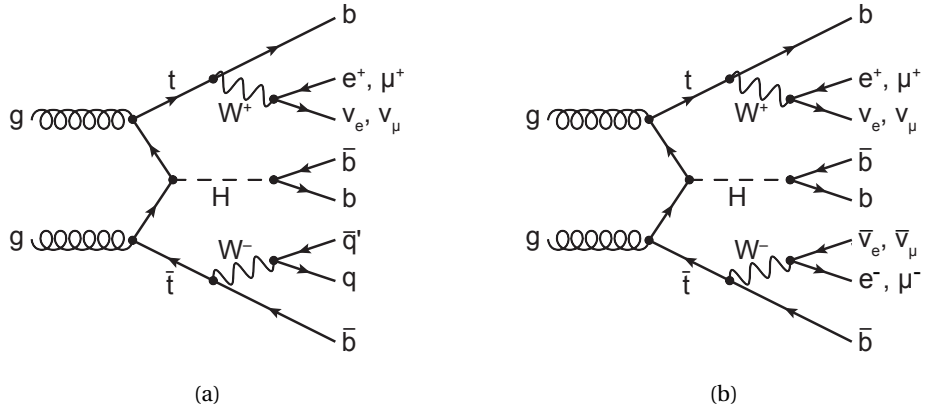


Figure 5.1: Tree-level Feynman diagrams for the production of the Higgs boson in association with a top quark pair and subsequent decay of the Higgs boson to a bottom quark pair. The single-lepton decay of the $t\bar{t}$ system is shown in (a) and the dileptonic decay in (b).

Depending on the transverse momenta of the objects in the $t\bar{t}H$ process, events can be further classified into the *resolved* and *boosted* phase space regions. The resolved events are so named because we can *resolve* all small jets individually since the events contain low-p_T objects. The boosted events contain a boosted Higgs boson and hadronically decaying top quark, which have collimated decay products produced too close together to individually resolve with small jets. For these boosted objects we make use of large jets to capture their decays. The dilepton channel in this analysis has a resolved phase space only, whereas the semileptonic channel has both a resolved and a boosted contribution. The boosted dileptonic cross-section

is too small to analyse with the current dataset but this channel could be added in the future. The boosted semileptonic category was added for the first time in this round of the analysis.

To obtain the final analysis results, all regions from the dilepton resolved, semilepton resolved, and semilepton boosted categories are combined. These regions include signal and control regions which are defined according to the number of leptons, (large) jets, and b-jets in the events. The analysis is optimised by using multivariate techniques to distinguish signal from background events. An overview of the full analysis strategy is shown in figure 5.2 and each step is explained in detail in the following sections.

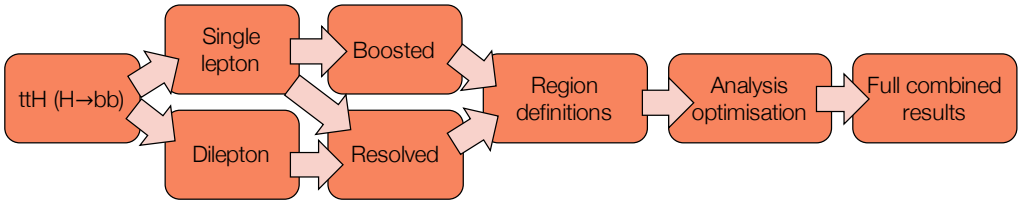


Figure 5.2: Overview of the $t\bar{t}H(H \to b\bar{b})$ analysis strategy.

5.3 Objects and event selection

The analysis makes use of jets and leptons in order to select the events of interest. We also rely heavily on the b-tagging of jets for our region definitions. The basic requirements for reconstructing leptons and jets in the analysis are summarised here together with the requirements for our event selection.

5.3.1 Data and triggering

The analysis uses 36.1 fb^{-1} of pp collision data at a centre-of-mass energy of $\sqrt{s} = 13$ TeV taken with the ATLAS detector in 2015 (3.2±0.1 fb^{-1}) and 2016 (32.9±0.7 fb^{-1}). We only use events that have at least one vertex associated with two or more tracks with $p_T > 0.4$ GeV. The level of pile-up in this dataset ranges from 8 to 45 interactions per bunch crossing (see figure 2.3). In order to separate the hard scatter vertex of interest from pile-up vertices, we label the primary vertex as the one with the largest sum of the squares of the transverse momenta of associated tracks. The events selected for this analysis are recorded using single-lepton triggers with requirements as shown in table 5.1. The lepton identification criteria and isolation requirements are detailed in reference [68] for electrons and in reference [70] for muons. The identification requirements are defined at three operating points: loose, medium, and tight. A fourth operating point named *gradient* exists for the isolation requirements; this working point becomes more stringent as the lepton p_T decreases.

Object	p_T threshold [GeV]	Identification criterion	Isolation requirement
Electrons	24 (26)	Medium (Tight)	Gradient
	60	Medium	None
	120 (140)	Loose	None
Muons	20 (26)	Loose (Medium)	Loose (Gradient)
	50	None	None

Table 5.1: Single-lepton triggers used for the analysis. The parameters are given for 2015 data and, if different, for 2016 data in brackets.

5.3.2 Leptons

As explained in section 3.2, the reconstruction of leptons relies on ID tracking and additionally the EM calorimeter for electrons, the muon spectrometer for muons, and both the EM and HAD calorimeter systems for taus. For an object to be reconstructed as a lepton, all lepton tracks must match the main primary vertex of the event. The full lepton reconstruction requirements are shown in table 5.2. The isolation requirements are set to reduce the contribution of non-prompt leptons coming from hadronic decays. The identification criteria for electrons, muons, and taus are described in references [68], [70], and [71] respectively. Any electron candidates in the calorimeter crack region ($1.375 \leq |\eta| < 1.52$) between the barrel and end-cap are excluded because a proper energy measurement is not possible there.

Lepton	Subdetectors used	p_T cut	$\lvert\eta\rvert$ cut	Identification criterion	Isolation requirement
Electron	EM calorimeter energy deposits matched to ID tracks	10 GeV	< 2.47	Loose	Loose
Muon	(Partial) muon spectrometer tracks matched to ID tracks	10 GeV	< 2.50	None	Loose
Tau	Full calorimeter energy deposits matched to ID tracks	25 GeV	< 2.50	Medium	None

Table 5.2: The requirements for the reconstruction of leptons in the analysis. In addition to these requirements, all lepton tracks must match the main primary vertex of the event.

5.3.3 Jets

The $t\bar{t}H(H \to b\bar{b})$ analysis uses both small and large jets, as defined in sections 4.3 and 4.4, which are implemented in the FastJet package [81]. The small jets are constructed from topo-clusters calibrated to the EM scale and clustered with the anti-k_T algorithm (see section 4.2) with a radius parameter of $R = 0.4$. They are calibrated to the jet energy scale as described in section 4.3.1. After the calibration, the jets are required to have $p_T > 25$ GeV and $|\eta| < 2.5$. The JVT discriminant (see section 4.3.3) of all jets is required to be > 0.59 in order to reject jets

originating from pile-up. All jets need to pass the loose jet cleaning selection (see section 4.3.2).

Jets are tagged as containing b-hadrons by use of the MV2c10 algorithm discussed in section 4.3.4. The MV2c10 BDT output is used as a binned discriminant for all jets. We define five bins corresponding to the four WPs (see table 4.1) and an extra bin for jets that fail the loose WP. The MV2c10 value is checked for each jet, and the jet is assigned an integer value according to the WP that it passes. If a jet passes the very tight WP, it is assigned a b-tagging score of 4, whereas it is assigned a 0 if it fails the loose WP. This binned b-tagging method is referred to as *pseudo-continuous b-tagging* and is detailed in table 5.3.

WP	***b*-jet efficiency**	**Pseudo-continuous *b*-tagging score**
Very tight	60%	4
Tight	70%	3
Medium	77%	2
Loose	85%	1
None	n.a.	0

Table 5.3: The definition of the pseudo-continuous b-tagging method used in the analysis.

We will discuss both trimmed and reclustered large jets, as described in section 4.4. The large jets have a radius parameter of $R = 1.0$ as decided upon by the studies shown in section 4.5. The topoclusters used for the standard large jets are calibrated with the LCW hadronic calibration (see section 4.1). Since reclustered large jets are built from the small jets, they indirectly consist of topoclusters at the EM scale. All large jets are required to have a jet mass of at least 50 GeV, a p_T between 200 and 1500 GeV, and $|\eta| < 2$. The reclustered jets are required to have at least two subjets and the standard jets are trimmed with $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$.

5.3.4 Overlap removal

Since both leptons and jets are constructed from calorimeter energy deposits, an overlap removal procedure is applied to avoid double counting of these deposits. The six steps involved in this procedure are carried out in the following order:

1. Remove muons if they are within $\Delta R < 0.4$ from the nearest jet with ≥ 3 associated tracks.
2. Remove closest jet to a muon within $\Delta R < 0.4$ if the jet has < 3 associated tracks.
3. Remove the closest jet within $\Delta R < 0.2$ from the electron.
4. Remove electrons if, after step 3, there is a jet within $\Delta R < 0.4$ of the electron.
5. Remove large jets that overlap with the primary electron in the event.
6. Remove a τ candidate if it is within $\Delta R < 0.2$ of an electron or muon.

5.3.5 Event selection

In the semilepton channel, events must contain exactly one reconstructed lepton with $p_T > 27$ GeV and no other leptons that pass the requirements as specified in table 5.2. The selected lepton should be within $\Delta R < 0.1$ of a lepton with the same flavour reconstructed by the trigger algorithm. Events in the dilepton channel are required to have exactly two reconstructed leptons of opposite electric charge, the leading of which needs to have $p_T > 27$ GeV and the subleading a p_T of at least 15 GeV in the *ee* channel or 10 GeV in the $e\mu$ and $\mu\mu$ channels. In addition to these p_T requirements, all selected electrons need to pass the *tight* identification criterion [68], and the selected muons the *medium* criterion [70]. All selected leptons need to satisfy the *gradient* isolation requirement.

In order to remove any overlap with other $t\bar{t}H$ searches [105], events are vetoed if they contain one or more τ lepton candidates in the dilepton channel and two or more τ candidates in the singe-lepton channel. The τ leptons are required to have $p_T > 25$ GeV, $|\eta| < 2.5$ and pass the *medium* τ identification criterion [71]. Events are also vetoed if they have at least one fake jet identified in the jet cleaning procedure described in section 4.3.2.

5.4 Signal and background modelling

The nominal MC samples for the $t\bar{t}H(H \to b\bar{b})$ analysis are all generated with the fullsim procedure (see section 3.1.2) to achieve the best precision of data modelling. The fastsim is used for some of the alternative samples which are used for the systematic uncertainties (see section 7.3). An overview of the event generators used for the $t\bar{t}H$ signal process and its main background, $t\bar{t}$+jets, is shown in table 5.4. Pile-up effects are simulated by overlaying inelastic collisions generated with PYTHIA8.186 [51] onto the hard scatter events. The generation of the samples for this analysis will be discussed in more detail in this section.

An overview of the cross-section measurements carried out by ATLAS of the $t\bar{t}H$ signal and several of the background processes is shown in figure 5.3. The $t\bar{t}H$ signal has a very small cross-section of 0.507^{+35}_{-50} pb compared to the background processes; the main background in the analysis is $t\bar{t}$+jets which has a cross-section of 832^{+46}_{-51} pb. This is one of the main challenges of the analysis.

5.4.1 $t\bar{t}H$ signal

The ME of the $t\bar{t}H$ signal is modelled using MADGRAPH5_aMC@NLO (MG5_aMC@NLO) [43] version 2.3.2 which provides NLO accuracy. The NNPDF3.0NLO parton distribution function [39] is used. The ME calculation is interfaced to PYTHIA8.2 [51] for the parton shower and hadronisation model, using the A14 ATLAS tune to data [107]. The Higgs boson mass is fixed to 125 GeV and all Higgs boson decay modes are considered using the latest branching ratio calculations from reference [108].

Process	Event generator	Usage	Details
$t\bar{t}H$	MG5 + PYTHIA8	Nominal sample	
$t\bar{t}H$	MG5 + HERWIG++	Evaluate uncertainty on choice of PS and hadronisation model	
$t\bar{t}$+jets	POWHEG+ PYTHIA8	Nominal sample	$t\bar{t}+\geq 1b$ component reweighted to SHERPA4F sample
$t\bar{t}$+jets	POWHEG+ PYTHIA8	Additional statistics for boosted analysis	Filtered sample for high-p_T phase space
$t\bar{t}$+jets	POWHEG+ PYTHIA8	Evaluate uncertainty on radiation model	$t\bar{t}+\geq 1b$ component reweighted to SHERPA4F sample and variations on μ_R, μ_F, h_{damp} and A14 Var3c parameters
$t\bar{t}$+jets	SHERPA5F	Evaluate uncertainty on choice of ME calculation	$t\bar{t}+\geq 1b$ component reweighted to SHERPA4F sample
$t\bar{t}$+jets	POWHEG+ HERWIG7	Evaluate uncertainty on choice of PS and hadronisation model	$t\bar{t}+\geq 1b$ component reweighted to SHERPA4F sample
$t\bar{t}+\geq 1b$	SHERPA4F	Evaluate uncertainty of 4FS vs. 5FS generator	$t\bar{t}+\geq 1b$ component reweighted to SHERPA4F sample
$t\bar{t}+\geq 1c$	MG5 + HERWIG++	Evaluate uncertainty of 3FS vs. 5FS generator	

Table 5.4: Summary of the event generators used for the $t\bar{t}H$ signal and the main background $t\bar{t}$+ jets. For each of the alternative generators used for $t\bar{t}$+ jets, the fractions of $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$, and $t\bar{t}$+ light are reweighted to the nominal POWHEG+ PYTHIA8 sample. MG5 stands for the MADGRAPH5_aMC@NLO generator and FS refers to the flavour scheme used in the generator, as discussed in section 3.1.1.

5.4.2 $t\bar{t}$+jets background

The background to the $t\bar{t}H$ process is dominated by $t\bar{t}$+jets events. In the boosted single-lepton signal region, the fraction of background events coming from this process is 84% in simulated events. The top quark mass in the $t\bar{t}$ samples is set to 172.5 GeV. The ME of the nominal sample for this background is generated using POWHEG-BOX v2 [44, 45] with NLO accuracy, using the NNPDF3.0NLO PDF set. The h_{damp} parameter was set to 1.5 times the top quark mass as was optimised in reference [109]. This parameter sets an upper bound on the p_T of the first additional radiation to the $t\bar{t}$ system and controls the matrix element to parton shower matching in POWHEG. The parton shower and hadronisation is modelled with PYTHIA8.2 with the A14 tuning. The $t\bar{t}$ sample is normalised to the predicted cross-section calculated with the Top++2.0 program [110] at NNLO accuracy and next-to-next-to-leading logarithmic (NNLL) resummation of soft gluon terms: 832^{+46}_{-51} pb [20–23].

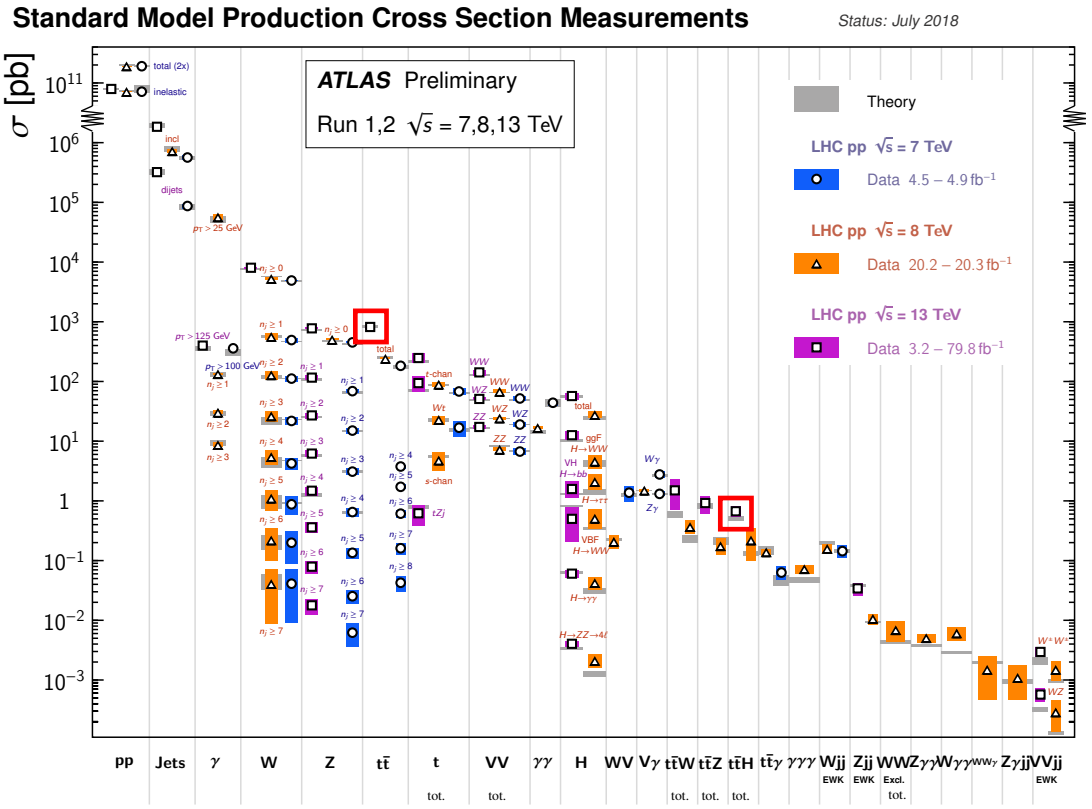


Figure 5.3: Summary of several SM total and fiducial production cross-section measurements made by ATLAS [106]. The measurements for $t\bar{t}H$ and $t\bar{t}$ at $\sqrt{s} = 13$ TeV are highlighted with the red boxes.

Flavour classification

The $t\bar{t}$+jets background is divided into three categories according to the flavour of jets not originating from the decay of the $t\bar{t}$ system (also called additional jets) in the event, similar to the procedure carried out in [25]. The anti-k_T jets with $R = 0.4$ are flavour-labelled at MC truth level according to the number of b- or c-hadrons within $\Delta R < 0.4$ of the jet axis. All truth jets are required to have $p_T > 15$ GeV and $|\eta| < 2.5$. A jet is labelled as b-jet if it is matched to exactly one b-hadron (with $p_T > 5$ GeV) and as B-jet if it is matched to two or more b-hadrons. In the same way we define c-jets and C-jets. An event that has at least one b-jet or B-jet is labelled as $t\bar{t}+\geq 1b$. Any events that are not already categorised as $t\bar{t}+\geq 1b$ and contain at least one c-jet or C-jet are labelled $t\bar{t}+\geq 1c$. In this classification, we only consider jets that do not originate from top quark or W boson decays. The $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ events are jointly referred to as $t\bar{t}+$ heavy flavour (HF). The remaining events that do not contain any HF jets are labelled as $t\bar{t}+$ light.

The $t\bar{t}+\geq 1b$ contribution is of particular importance because it represents the largest fraction of our total background. This contribution is subject to an even finer categorisation for the purpose of comparisons to other $t\bar{t}$+jets event generators, to assign correction factors, and to estimate uncertainties. Events are labelled $t\bar{t}+b$ if they contain exactly one single truth b-jet and as $t\bar{t}+b\bar{b}$ if they contain exactly two truth b-jets. The events labelled as $t\bar{t}+B$ contain

exactly one B-jet and the rest is labelled as $t\bar{t}+\geq 3b$. A fifth subcategory, $t\bar{t}+b$ (MPI/FSR), is defined for any events with additional b-jets originating from MPI or FSR.

In order to improve the $t\bar{t}+\geq 1b$ modelling we reweight the relative contributions of each of the $t\bar{t}+\geq 1b$ subcategories from POWHEG+PYTHIA8 to the SHERPA+OPENLOOPS [46, 49] prediction. This SHERPA+OPENLOOPS sample is a state-of-the-art dedicated NLO $t\bar{t}+\geq 1b$ sample including parton showering and hadronisation with massive b-quarks [111] (this is the 4F scheme, see section 3.1.1). Since this is the most precise MC prediction for the $t\bar{t}+\geq 1b$ process available at present, it is expected to give a more accurate $t\bar{t}+\geq 1b$ modelling than our nominal POWHEG+PYTHIA8 sample. At present, we cannot use the SHERPA+OPENLOOPS sample directly as our nominal $t\bar{t}+\geq 1b$ sample because event generation takes too long and there exists no clear prescription on how to combine the dedicated $t\bar{t}+\geq 1b$ sample with the inclusive $t\bar{t}$+jets samples and remove the overlap between them.

The SHERPA+OPENLOOPS sample uses the CT10 4F scheme PDF set [41, 42] which implements massive b-quarks. The nominal POWHEG+PYTHIA8 sample is calculated in the 5F scheme in which the b-quarks are treated the same as other massless partons. In the 5F scheme, the b-quarks are included in the initial state, whereas the 4F scheme includes them in the final state. The SHERPA+OPENLOOPS sample using the 4F scheme will be called SHERPA4F from here onward. A comparison of the predicted fractions of the $t\bar{t}+\geq 1b$ subcategories between the two generators is shown in figure 5.4. The $t\bar{t}+b$ (MPI/FSR) subcategory is not present in the SHERPA4F sample and these events are thus not reweighted.

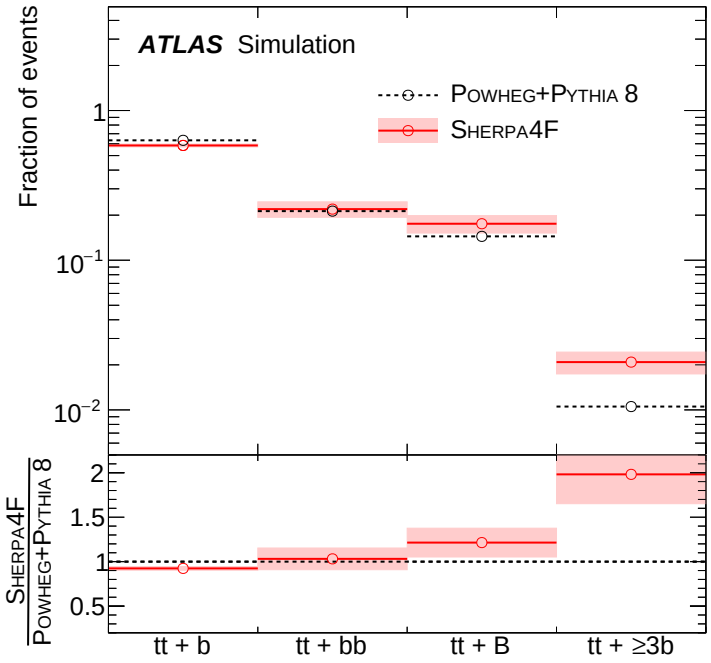


Figure 5.4: Comparison of the POWHEG+PYTHIA8 to SHERPA4F prediction of relative fractions of $t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$, and $t\bar{t}+\geq 3b$ subcategories [103]. The uncertainty bands on the SHERPA4F prediction consist of several sources which are detailed in section 7.3.2.

Boosted filtered samples

In addition to the nominal $t\bar{t}$ samples, we use two $t\bar{t}$ samples that are filtered in order to select events within the boosted phase space. This ensures a good statistical coverage in the high-p_T region that we probe. There is a c-filtered sample which contains at least one additional jet initiated by a c-quark, and a b-filtered sample with at least one additional b-quark initiated jet. The event selection in both samples requires the hadronic top quark to have $p_T > 200$ GeV and/or the $t\bar{t}$ system to have $p_T > 150$ GeV. The extra boosted filtered samples are only used in the BDT training for the boosted signal region (SR) (see section 5.7) and are not included in any of the subsequent fitting procedures and results.

The boosted filtered samples are combined with the nominal $t\bar{t}$ samples and the overlap between them is removed. The nominal $t\bar{t}$ samples used in the analysis represent 151 fb^{-1} and 442 fb^{-1} for the inclusive and b-filtered samples, respectively. The boosted filtered samples represent 6218 fb^{-1} for the b-filtered sample and 1281 fb^{-1} for the c-filtered sample. A comparison between the $t\bar{t}$ event yields for the nominal samples and when including the boosted filtered samples is shown in table 5.5. A large statistical enhancement is seen for the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ contributions. Figure 5.5 shows the truth p_T distributions of the $t\bar{t}$ system and the hadronic top quark with the nominal and boosted filtered samples compared. The distributions are very similar and the uncertainties at high p_T are reduced when including the boosted filtered samples.

Sample	Nominal		Including filtered samples		
	Raw	Weighted	Raw	Weighted	Change in stats
$t\bar{t}+\geq 1b$	1971	219 ± 6	19780	206 ± 4	$\times 10$
$t\bar{t}+\geq 1c$	438	149 ± 9	2622	144 ± 6	$\times 6$
$t\bar{t}+$ light	280	119 ± 10	661	118 ± 9	$\times 2$

Table 5.5: The yields for the three different $t\bar{t}$ contributions with the nominal $t\bar{t}$ samples (left) and when including the boosted filtered $t\bar{t}$ samples (right).

5.4.3 Other real backgrounds

On top of the $t\bar{t}$+jets background, the $t\bar{t}H(H \to b\bar{b})$ signal region has a small component of $t\bar{t}V$ ($t\bar{t}W$ and $t\bar{t}Z$) backgrounds (2% in the boosted signal region) and a few non-$t\bar{t}$ backgrounds (14% in the boosted signal region). We include the small contribution from Higgs boson production in association with a single top quark as background. Samples of single top quarks produced with a W boson and a Higgs, tWH, and with additional jets, $tHqb$, are considered. The other Higgs production modes were found to be negligible and are not included.

- $t\bar{t}W$ and $t\bar{t}Z$: Matrix element MG5_aMC@NLO interfaced to PYTHIA8.2 with the NNPDF3.0NLO PDF set and A14 tuning for parton shower and hadronisation.
- Wt and s-channel single top quark: Matrix element POWHEG-BOX v1 at NLO accuracy using the CT10 PDF set interfaced to PYTHIA6.4 [50] with the Perugia 2012 tuning [112].

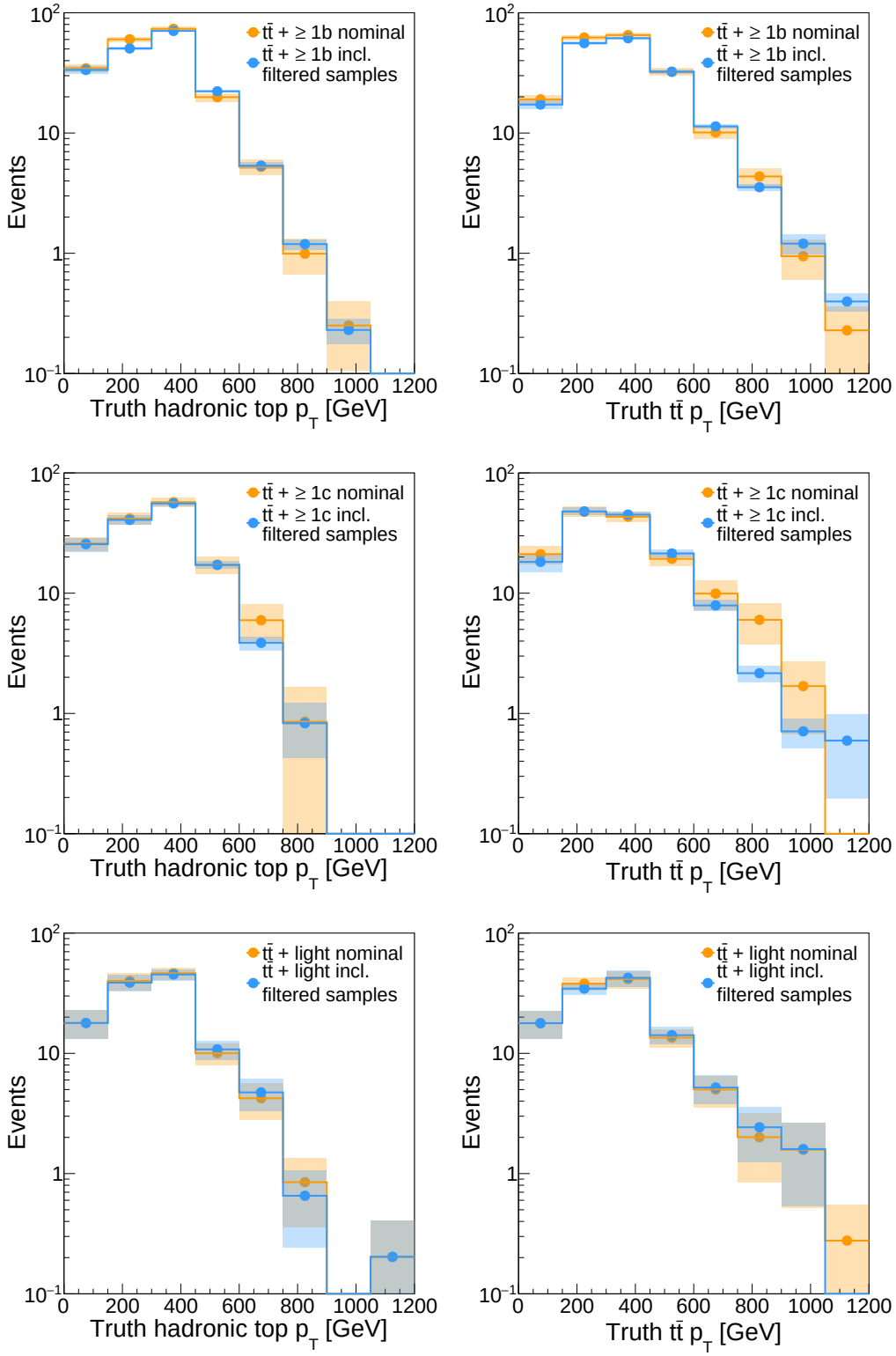


Figure 5.5: The truth p_T distributions of the hadronic top quark (left column) and $t\bar{t}$ system (right column) in the boosted SR. A comparison is made between the nominal $t\bar{t}$ MC samples (orange) and the full sample collection including boosted filtered samples (blue). The top row shows $t\bar{t}+\geq 1b$ events, the middle row shows $t\bar{t}+\geq 1c$, and the bottom row shows $t\bar{t}+$ light.

- t-channel single top quark: Matrix element POWHEG-BOX v1 at NLO accuracy using four-flavour PDF set CT10 4F which accounts for massive b-quarks. The parton shower and hadronisation are simulated with PYTHIA6.4 using the Perugia 2012 tuning.
- W+jets, Z+jets, and diboson+jets: SHERPA 2.2.1 for matrix element and parton shower. The NNPDF3.0NLO PDF set is used in combination with a dedicated parton shower tuning. The Z+heavy flavour jets contribution is scaled up by a factor 1.3 which is extracted from dedicated control regions in data.
- $t\bar{t}WW$, tZW, and 4-top: MG5_aMC@NLO interfaced to PYTHIA8 with the A14 tune and NNPDF3.0NLO PDF set.
- tZ: MG5_aMC@NLO interfaced to PYTHIA6 with the Perugia 2012 tune.
- tWH: Matrix element MG5_aMC@NLO interfaced to HERWIG++ [53] with the CTEQ6L1 PDF set [40].
- $tHjb$: Matrix element MG5_aMC@NLO at LO accuracy, interfaced to PYTHIA8 using the CT10 4F PDF set.

5.4.4 Fake lepton backgrounds

An additional category of background events is considered which constitutes about 4% of the total background in the boosted signal region; the *fakes*. These are jets or photons misidentified as a lepton, or non-prompt leptons originating from the decays of long-lived particles. The main source of fake electrons is caused by jets with large deposits in the EM calorimeter and fake muons mostly originate from in-flight hadron decays. In the single-lepton channel, fakes originate mostly from QCD multijet events, whereas the fakes in the dilepton channel come from other sources such as $t\bar{t}$ single-lepton. Because the QCD multijet processes are hard to model theoretically, we use recorded data to estimate their contributions in the single-lepton channel, instead of simulated MC events as for the other backgrounds.

We use the matrix method which calculates an event weight to be applied to data in order to model the fake contribution [113]. A control region is selected and the efficiencies of fakes and real leptons passing the loose and tight lepton identification requirements is measured. Events passing the tight lepton requirement are a subset of the events passing the loose requirement. We can write the total number of events passing the loose selection as

$$N^{\text{loose}} = N_r^{\text{loose}} + N_f^{\text{loose}}, \tag{5.1}$$

where N_r refers to events with a real lepton and N_f to events with a fake lepton. The total number of events passing the tight selection can then be written as

$$N^{\text{tight}} = N_r^{\text{tight}} + N_f^{\text{tight}} = \epsilon_r N_r^{\text{loose}} + \epsilon_f N_f^{\text{loose}}, \tag{5.2}$$

where ϵ_r (ϵ_f) is the fraction of events in the loose selection containing a real (fake) lepton which also passes the tight selection requirement. We can combine these two expressions to get the number of events passing the tight selection and containing a fake lepton as

$$N_f^{\text{tight}} = \frac{\epsilon_f}{\epsilon_r - \epsilon_f}\left(\epsilon_r N^{\text{loose}} - N^{\text{tight}}\right). \tag{5.3}$$

The efficiencies ϵ_r and ϵ_f are measured in control regions enriched in real leptons and fake leptons, respectively. They are expected to be dependent on kinematic variables characterising the events in these regions and are therefore parametrised as a function of different sets of variables for the resolved and boosted analyses. The event weight w_i applied to data is given as a function of the event kinematics k_i by

$$w_i = \frac{\epsilon_f(k_i)}{\epsilon_r(k_i) - \epsilon_f(k_i)} \left(\epsilon_r(k_i) - \delta_i\right), \tag{5.4}$$

where δ_i is one for events passing both the tight and loose requirements, and zero for events passing only the loose requirement. The final fakes background estimation is given by the sum of w_i over all events.

The fakes in the dilepton channel are estimated by simulated MC samples which do not contain two opposite-sign leptons. In order to improve the prediction, the background is normalised to data in a control region with two same-sign leptons.

5.5 Boosted signal region optimisation

The boosted category is combined with the resolved single-lepton and dilepton channels in the final results. The boosted regime accesses different kinematics than the resolved channel which gives us access to a different phase space where we can use innovative and interesting techniques. The addition of a boosted channel can also improve the sensitivity of the full analysis because it accesses extra information and has a simplified combinatorial background due to multiple decay products being caught in one large jet. Since this is the first time that a boosted region is included in the $t\bar{t}H(H \to b\bar{b})$ analysis, the event selection was designed from scratch. In the designing of our signal region, we have followed the process as depicted in figure 5.6.

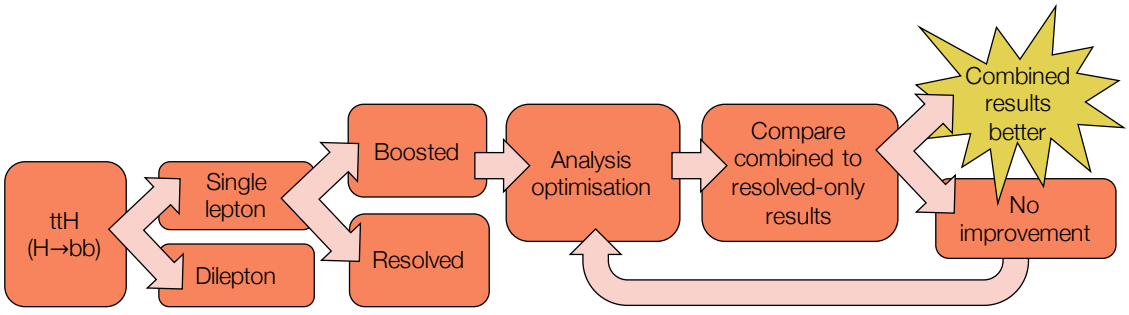


Figure 5.6: Overview of the boosted $t\bar{t}H$ analysis strategy.

We have explored several options and methods in the boosted channel to optimise the analysis. Every option was put to the test by making a comparison between the combined final results where we include the boosted signal region, and the final results using only the resolved regions. We aim for an improvement of the analysis sensitivity when including the boosted channel. If we see no improvement of the sensitivity, or indeed a degradation, we go back to

the drawing board and try to improve our techniques. A degradation in the analysis sensitivity is possible because there is an overlap between the resolved and boosted events. These events are given to the boosted region and vetoed in the resolved analysis which could lead to a decrease in performance if the boosted analysis achieves less sensitivity.

Note that the choice of the boosted signal region was made before the final updates made to the resolved analysis, so the results here are compared to a baseline version of the resolved analysis established for the ICHEP conference in 2016 which differs from the final resolved analysis described later on.

5.5.1 The signal region options

Four candidate signal regions were investigated, tested, and their results compared. The signal regions under consideration are shown in figure 5.7 and described in detail below. The signal region d) using reclustered jets is my own design.

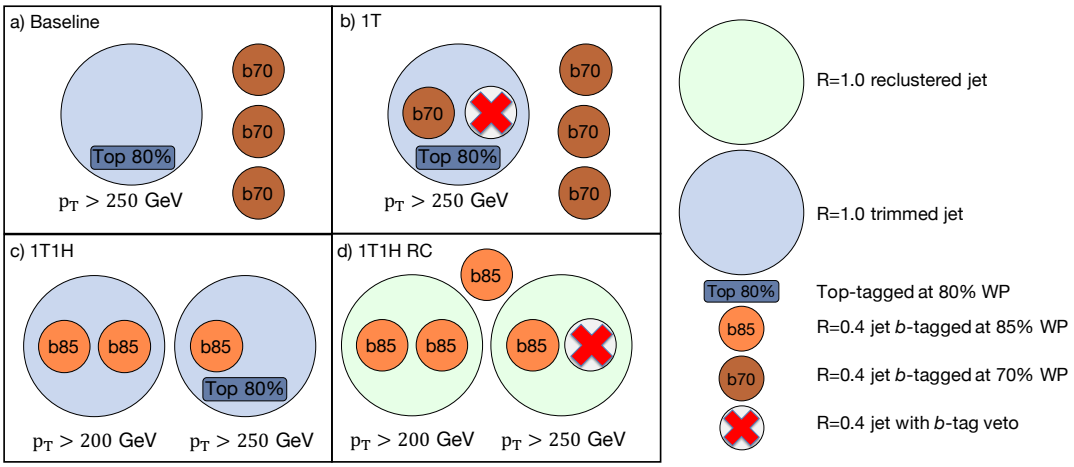


Figure 5.7: The four boosted signal regions under consideration.

a) Baseline: The baseline region is the benchmark to which the performance of the other three regions is compared. This selection targets events which contain at least one trimmed large jet which is top-tagged at the loose (80%) WP (see section 4.4.2), with $p_T > 250$ GeV. In addition to the large jet, this region requires at least three small jets which are b-tagged at the tight (70%) WP (see table 4.1). These three small jets need to pass the requirement that they are $\Delta R > 1.0$ away from all top-tagged large jets in the event.

b) 1T: Selects events like the baseline region described above, with the extra requirement that there is exactly one tight b-tagged small jet and at least one small jet which fails the tight b-tagging WP within $\Delta R < 1.0$ of the top-tagged large jet. This requirement improves the top-tagging purity.

c) 1T1H: This region is based on the presence of two large trimmed jets, one of which is the Higgs candidate jet which is identified first. The Higgs candidate jet is required to have $p_T > 200$ GeV and to be matched to at least two loose (85% WP) b-tagged small jets within

$\Delta R < 1.0$. In case there is more than one Higgs candidate in the event, we choose the one which has the highest sum of b-tag discriminant values of all small jets within $\Delta R < 1.0$ of the Higgs jet. The top candidate is a separate large jet with $p_T > 250$ GeV and is top-tagged at the loose (80%) WP. Additionally, the top candidate needs to have exactly one loose b-tagged small jet within $\Delta R < 1.0$ of its axis.

d) 1T1H RC: This region requires two large jets just like the 1T1H region, but the large jets here are reclustered jets instead of trimmed jets (see section 4.4.4). The Higgs candidate is identified as a large reclustered jet with $p_T > 200$ GeV and at least two subjets that are b-tagged at the loose WP. If there is more than one Higgs candidate in the event, we pick the one with the highest sum of b-tag discriminant values of all subjets. The top candidate is found by identifying another large jet with $p_T > 250$ GeV. The top is required to have exactly one subjet b-tagged at the loose WP and at least one subjet which fails the loose b-tagging WP. If there is more than one top candidate in the event, the one with the largest mass is chosen. Finally, events are required to have an additional small jet which is b-tagged at the loose WP and is not a subjet of either the Higgs or top candidate reclustered jets.

The 1T1H RC region does not use the dedicated top tagger because this algorithm makes use of the N-subjettiness variable τ_{32}. This variable needs to be computed directly from the topocluster jet constituents, as shown in equation 4.8. This could be done for a reclustered jet by taking the jet constituents of its subjets, but this means that we no longer can simply use the small jet uncertainties for our reclustered jets and would need a dedicated set of uncertainties for the energy clusters. These uncertainties were not defined at the time of this analysis and it is not known how large of an impact these would have on the total uncertainty and analysis sensitivity.

5.5.2 Composition of the signal regions

Each of the signal regions differ in terms of signal purity, background composition, and selected event kinematics. The expected signal and background yields are shown in table 5.6 along with the S/B and S/$\sqrt{\text{B}}$ values which give an indication of the sensitivity that these regions can achieve. The 1T region has the largest values of these figures of merit, whereas the 1T1H and 1T1H RC regions have a similar, lower, purity. The background composition of each of the regions is shown in figure 5.8. The 1T region has by far the largest contribution of $t\bar{t}+\geq 1b$ background because of the tight b-tagging WP used. The 1T1H and 1T1H RC have larger contributions from $t\bar{t}+\geq 1c$ and $t\bar{t}+$ light due to the use of the loose b-tagging WP. In general, we would prefer a region with a lower contribution in $t\bar{t}+\geq 1c$ and $t\bar{t}+$ light because they are less well understood than the $t\bar{t}+\geq 1b$ background. The $t\bar{t}+\geq 1b$ has a dedicated NLO MC sample whereas this is not available for the others.

	Baseline	1T	1T1H	1T1H RC
Signal (S)	30	16	27	15
Total background (B)	875	234	1065	370
S/B	0.03	0.07	0.03	0.04
S/$\sqrt{B}$	1.0	1.1	0.8	0.8

Table 5.6: Expected signal and background yields using 36.1 fb^{-1} of data for each potential signal region.

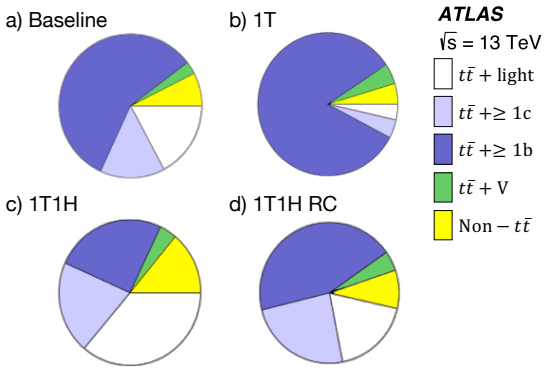


Figure 5.8: The background composition of the four boosted signal regions under consideration.

5.5.3 Overlap with resolved channel

Since the events in the boosted signal region have to be vetoed in the resolved single-lepton analysis, the amount of overlap between these analyses is of importance. Removing many events from the resolved analysis will degrade its sensitivity. This sensitivity can be regained, or even surpassed, by the boosted analysis, depending on its final performance. Since the boosted channel is added for the first time and we are not yet sure of its final sensitivity, we strive for a boosted signal region selection with the smallest possible overlap with the resolved regions. Since the final resolved region definitions were not fixed yet when the boosted signal region was decided upon, we are basing these results on the resolved regions' definition from the previous analysis round. These regions are defined by the number of small jets and number of jets b-tagged at 70% WP in the event. The minimum number of jets is four and the minimum number of b-jets is two.

The nine semileptonic resolved signal and control regions are shown in table 5.7 alongside the corresponding overlap of signal events with each of the boosted signal regions under consideration. The baseline region is left out since it is not an actual candidate for the final analysis. Events in the final column do not fall in any of the resolved regions and are thus additional events that would otherwise not be included in the analysis. The three red columns indicate the resolved signal regions, in which minimisation of the overlap is most important. We see that the 1T region has the largest overall overlap with the resolved signal regions. It has an especially large overlap ($\sim$ 13%) in the $\geq$ 6j,$\geq$4b region which is the purest, most sensitive signal region of the resolved analysis. The 1T1H and 1T1H RC regions have a similar level of overlap and both have some events that are not considered in the resolved analysis at all.

	4j,2b	4j,3b	4j,≥4b	5j,2b	5j,3b	5j,≥4b	≥6j,2b	≥6j,3b	≥6j,≥4b	None
Resolved	159	59.9	7.6	256	127	29.8	554	319	117	-
1T	-	-	-	-	-	1.5	-	-	15	-
1T1H	0.31	0.37	0.04	1.2	1.6	0.56	5.7	9.0	5.5	2.9
1T1H RC	-	-	-	0.21	0.45	0.48	2.5	5.9	4.9	0.56

Table 5.7: Number of signal events expected in each of the resolved regions for the resolved selection and each of the potential boosted signal regions. The regions are defined by the number of small jets, *j*, and *b*-jets at the tight working point, *b*, in the events. The three columns in red indicate the resolved signal regions.

5.5.4 Expected limits

In order to judge how well each of the boosted regions performs, we need to check how they perform in the final limit set on the parameter of interest for this analysis, $\mu_{t\bar{t}H}$, which is the ratio of the observed $t\bar{t}H$ cross-section over the expected SM result. The full fitting and limit setting procedure will be explained in chapter 6, however for now it is sufficient to understand that we aim to make the limit on $\mu_{t\bar{t}H}$ and the errors on this parameter as low as possible. The expected limits for a fit including only statistical uncertainties, and including the full systematics, are shown in table 5.8. These fits are performed on Asimov data (see section 6.3). The fits performed here show the combination of all resolved single-lepton regions and the boosted region combined. The last column shows the expected error on the parameter of interest. We see that all of the boosted regions achieve a small improvement over the resolved-only analysis, and the 1T1H RC region outperforms the others for both of the limits and the error bars.

	Statistics only limit	Full systematics limit	Error on $\mu_{t\bar{t}H}$
Resolved	0.607	1.01	$+0.57, -0.52$
1T combined	0.600	0.88	$+0.51, -0.46$
1T1H combined	0.606	0.87	$+0.51, -0.45$
1T1H RC combined	0.591	0.85	$+0.50, -0.43$

Table 5.8: Comparison of the statistics only and full systematics limits for the full single-lepton analysis using only the resolved regions, and when including each of the boosted signal regions under consideration. The error on the signal strength parameter $\mu_{t\bar{t}H}$ is also given.

5.5.5 Signal region selection

As seen from the previous sections, all signal regions have their pros and cons. The 1T region has the best S/B and S/$\sqrt{\text{B}}$ and the smallest contribution of $t\bar{t}+\geq 1c$ and $t\bar{t}+$ light background events which are less well understood than $t\bar{t}+\geq 1b$. However, this region does have the largest overlap with the resolved analysis, especially in its most sensitive signal region. The 1T1H still has a large overlap in the resolved ≥ 6j regions whereas the 1T1H RC region has a much smaller overlap which is desirable for the final combination. In general, the use of a signal region with two boosted objects is preferred over the 1T region since one of the reasons of performing this

analysis is to test out new techniques and get a better understanding of the boosted phase-space. Having a boosted Higgs as well as a boosted top means an overall more boosted region and is an opportunity to test out techniques for two distinct boosted objects.

The expected limits and error bars on $\mu_{t\bar{t}H}$ show that the 1T1H RC region is the most sensitive of the three options. This region also has the added benefit that it uses reclustered jets instead of trimmed jets, which means that there is no need for additional large jet uncertainties (see section 4.4.4) which decreases the overall systematic uncertainty of the combined result. This is also easier in practice since this extra set of uncertainties does not have to be calculated and added in the analysis software framework. Together with the fact that this region has the smallest overlap with the sensitive resolved regions, these reasons have lead the 1T1H RC region to be selected as the final boosted signal region to be used.

5.6 Event categorisation

The events selected for the $t\bar{t}H(H \to b\bar{b})$ analysis are categorised according to the number of (large) jets and b-jets in the events. Categories are labelled as *signal* regions when the $t\bar{t}H$ signal and $t\bar{t}+\geq 1b$ background are enhanced compared to other backgrounds. The remaining categories are labelled *control* regions and these provide constraints on backgrounds and systematic uncertainties in the combined fit.

5.6.1 Boosted region

The boosted analysis has one dedicated signal region which is the 1T1H RC region described in section 5.5.1. The events selected by the boosted region get priority over the resolved regions. There are no dedicated boosted control regions. Figure 5.9 shows the definition of the semileptonic boosted signal region which includes two reclustered large jets, one Higgs candidate and one top candidate, and at least five small jets of which at least four are b-tagged at the loose (85%) WP. The Higgs candidate is identified as the large reclustered jet with $p_T > 200$ GeV and ≥ 2 b-tagged subjets. This simple tagging strategy finds the correct Higgs jet in 47% of selected $t\bar{t}H$ events. This means that the Higgs candidate's subjets are truth matched within $\Delta R < 0.4$ to both b-jet daughters of the Higgs boson decay. In less than 0.5% of events, the top candidate jet contains the truth b-jets of the Higgs boson decay. In the remaining events, both b-jet daughters are not fully contained inside a single large jet.

The distributions of the truth level transverse momentum of the hadronically decaying top quark in $t\bar{t}H$ and $t\bar{t}$ events is shown in figures 5.10(a) and (b). The truth Higgs boson p_T is also shown in $t\bar{t}H$ events in figure 5.10(c). Distributions are shown for the boosted signal region selection and the purest resolved signal region containing at least 6 jets. These selections are made at reco jet level and the shown transverse momenta distributions are plotted at truth jet level. The boosted region clearly targets the higher p_T regions of these distributions compared to the resolved region.

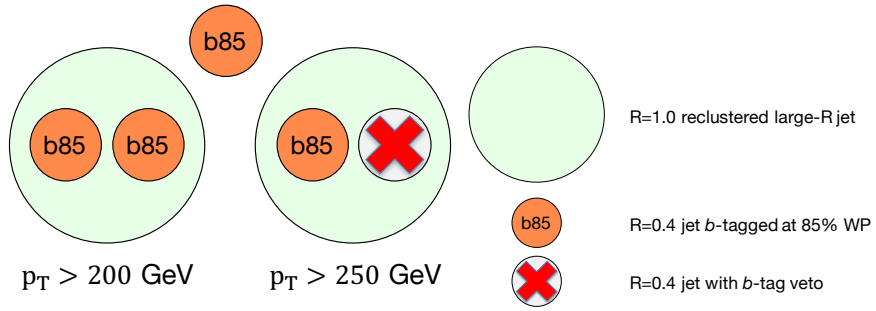
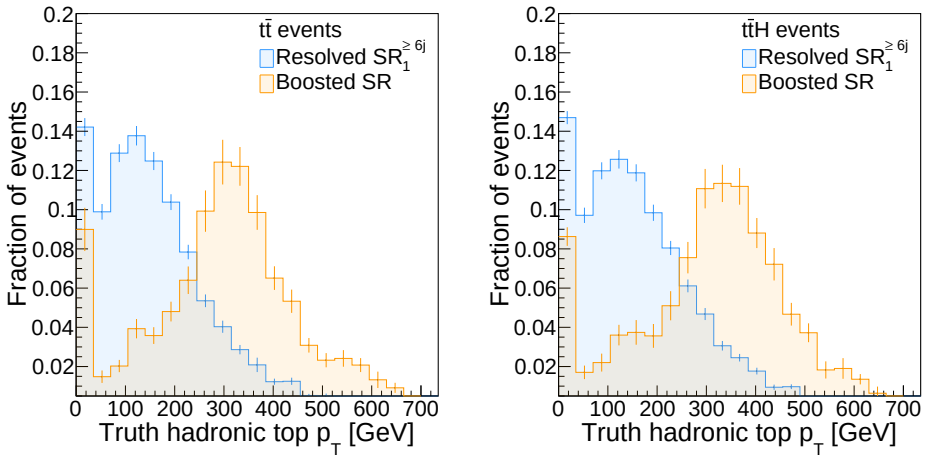
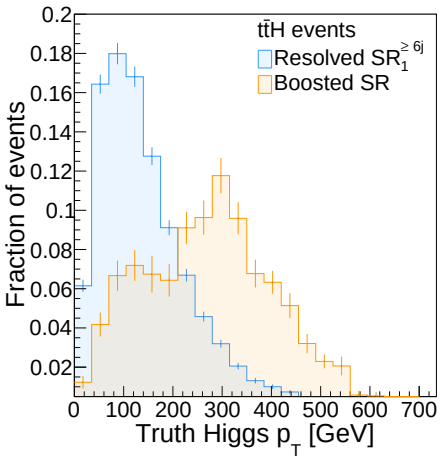


Figure 5.9: Definition of the semileptonic boosted signal region. The Higgs candidate jet has a p_T cut of 200 GeV and the top candidate jet a cut of $p_T > 250$ GeV.



(a)

(b)



(c)

Figure 5.10: The truth p_T distributions of (a) the hadronically decaying top in $t\bar{t}$ events, (b) the hadronically decaying top in $t\bar{t}H$ events, and (c) the Higgs boson in $t\bar{t}H$ events. The distributions are shown for the boosted signal region and for the purest resolved signal region with ≥ 6 jets.

Table 5.9 shows the boosted signal region event yields for signal and background events. Yields are shown pre-fit and post-fit, where the fit applied here is the full single-lepton fit with systematics. In this final version of the analysis, we have about 2.4% signal events. The $t\bar{t}H$ signal events consist for 85% of the $H \to b\bar{b}$ decay mode, and 15% of other Higgs decay modes. The background consists for 84% of $t\bar{t}$+ jets events, 40% of which is $t\bar{t}+\geq 1b$.

Sample	**Boosted SR**	
	Pre-fit	**Post-fit**
$t\bar{t}$ + light	177 ± 123	121 ± 36.9
$t\bar{t}$ + ≥1c	168 ± 69.9	224 ± 39.9
$t\bar{t}$ + ≥1b	236 ± 88.9	248 ± 43.1
$t\bar{t}+W$	5.41 ± 1.44	5.31 ± 1.19
$t\bar{t}+Z$	10.7 ± 2.20	10.3 ± 1.99
Wt channel	25.2 ± 18.7	24.2 ± 17.2
t channel	1.15 ± 1.52	1.22 ± 1.56
Other top sources	5.22 ± 2.11	5.06 ± 2.01
VV & V + jets	32.9 ± 15.2	32.2 ± 11.2
Fakes & NP (μ)	18.0 ± 12.2	20.1 ± 7.62
Fakes & NP (e)	11.0 ± 6.50	10.8 ± 6.35
tHjb	0.0748 ± 0.0231	0.0753 ± 0.0219
WtH	1.89 ± 0.271	1.89 ± 0.248
$t\bar{t}H$ ($H \to b\bar{b}$)	14.4 ± 1.62	9.51 ± 10.1
$t\bar{t}H$ ($H \to WW$)	1.15 ± 0.255	0.749 ± 0.823
$t\bar{t}H$ ($H \to$ other)	1.29 ± 0.476	0.873 ± 0.992
Total	710 ± 199	715 ± 41.0
Data	740	740

Table 5.9: Signal and background yields for the boosted SR, shown pre-fit and post-fit. The fit applied here is the full single-lepton fit with systematics. The uncertainty on the normalisations of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ is not defined pre-fit and therefore only included post-fit. For the $t\bar{t}H$ signal yield, the pre-fit values are the theoretical prediction and its corresponding uncertainty, whereas the post-fit values come from the signal strength measurement.

5.6.2 Resolved regions

The resolved event categorisation is much more involved than the boosted one because it selects many more events. The events are categorised into regions to separate signal events from reducible backgrounds. The resolved events in the single-lepton channel are first divided into those with exactly five jets and those with at least six jets. In the dilepton channel, the regions have either exactly three jets or at least four. A further subdivision is made according to the pseudo-continuous b-tagging scale detailed in table 5.3. All jets in each event are assigned an integer score from 0 to 4 according to the b-tagging WP that they pass. The four jets (three jets in the case of the dilepton regions with exactly three jets) with the highest b-tagging

score determine which region the event falls in. These regions are optimised in order to obtain categories enriched in either $t\bar{t}H$ and $t\bar{t}+b\bar{b}$, $t\bar{t}+b$, $t\bar{t}+\geq 1c$, or $t\bar{t}+$ light. Events in the first category make up the signal regions and contain $t\bar{t}+b\bar{b}$ as well as $t\bar{t}H$ since both processes have the exact same final state and $t\bar{t}+b\bar{b}$ is therefore an irreducible background to $t\bar{t}H$. The latter three categories make up the control regions and are used to control the modelling of the background processes.

Figure 5.11 shows the detailed definition of the signal and control regions for the semileptonic resolved channels. The regions in red are the signal regions and the ones in blue and white are the control regions. The resolved single-lepton channel has five signal regions, the purest of which is marked SR_1 in the ≥ 6 jets category shown in figure 5.11(b). In this region, there are four jets that each pass the *very tight* b-tagging WP (i.e. their b-tag discriminant score is 4). The second signal region, SR_2, in this figure contains events that have at least 6 jets of which the first three are b-tagged at the *very tight* WP and the fourth is b-tagged at either the *tight* or the *medium* WP. There is one more signal region with 6 or more jets, and there are also two signal regions with exactly 5 jets. The single-lepton analysis has six control regions: one for each of the $t\bar{t}+b$, $t\bar{t}+\geq 1c$, and $t\bar{t}+$ light backgrounds for both the 5 jets and ≥6 jets categories.

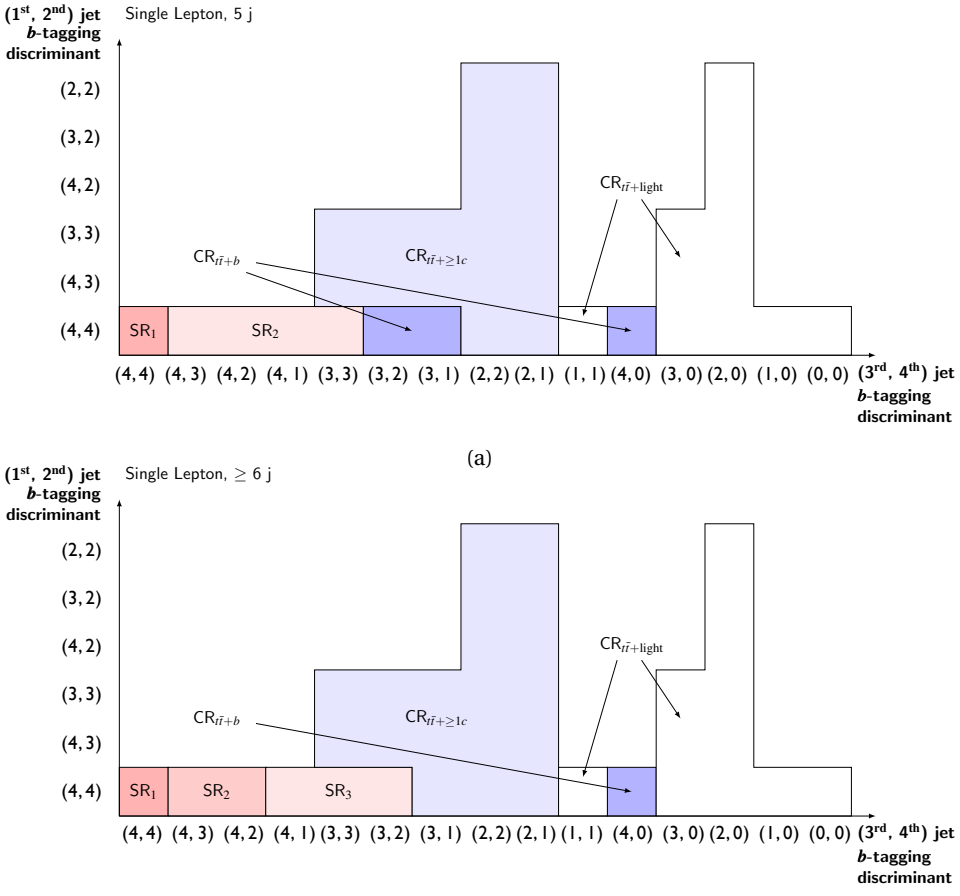


Figure 5.11: Definition of the regions for the semileptonic resolved channel, for events containing exactly 5 jets (a) and events containing at least 6 jets (b) [103]. The vertical axis shows the b-tagging discriminant value for the first and second jets whereas the horizontal axis shows this discriminant for the third and fourth jets. The jet ordering is based on the value of this discriminant in descending order.

The dilepton analysis has three signal regions which all fall into the ≥ 4 jets category. This channel has four control regions, two of which are in the ≥ 4 jets category (enriched in $t\bar{t}+\geq 1c$ and $t\bar{t}+$ light) and two in the 3 jets category (enriched in $t\bar{t}+\geq 1b$ and $t\bar{t}+$ light). The details about the dilepton region definitions are included in appendix A.1.

5.6.3 Composition of the regions

The fractional background contributions of each of the semileptonic control and signal regions can be seen in figure 5.12 (see appendix A.1 for the dilepton plots). As expected, the backgrounds in the signal regions are all dominated by $t\bar{t}+\geq 1b$. The $t\bar{t}H$ signal purity is shown in figure 5.13 where all $t\bar{t}H$ decays are counted as signal (see appendix A.1 for the dilepton plots). The decay to a pair of bottom quarks for which the analysis is designed represents 89%, 96%, and 86% of the total $t\bar{t}H$ signal events for the resolved dilepton, resolved single-lepton, and boosted single-lepton channels respectively. Each of the nine signal regions has its own level of signal purity, but all of them have an S/B value of at least 1.5%.

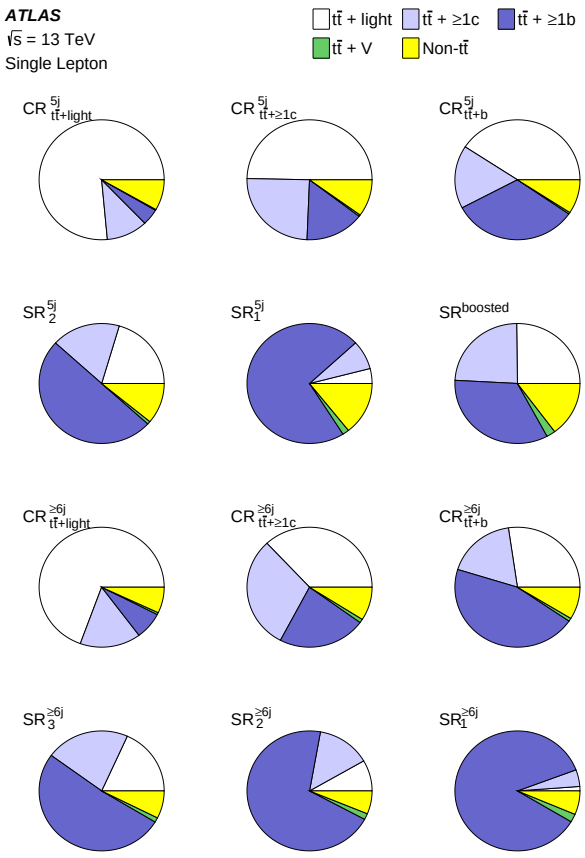


Figure 5.12: Background composition of the semileptonic signal and control regions including the boosted signal region [103]. The $t\bar{t}$ background is divided into $t\bar{t}+$ light, $t\bar{t}+\geq 1c$, $t\bar{t}+\geq 1b$, and $t\bar{t}+$ V contributions.

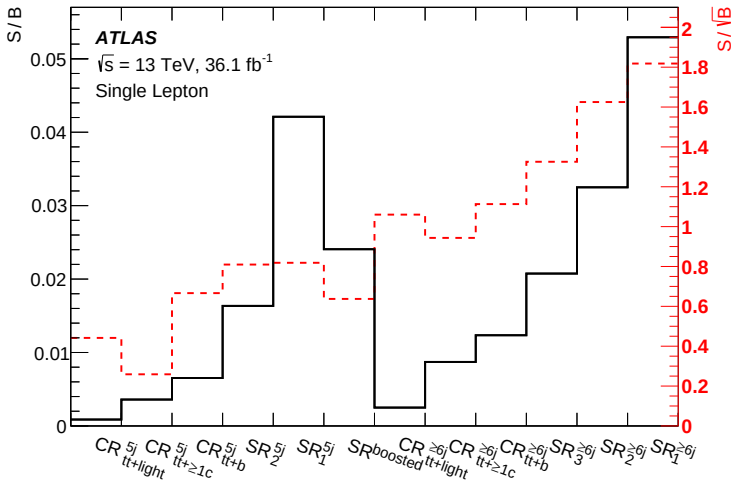


Figure 5.13: Purity of the semileptonic signal and control regions including the boosted signal region [103]. The S/B ratio is shown in black on the left vertical axis and the S/$\sqrt{B}$ is shown in red on the right vertical axis.

5.7 Multivariate analysis techniques

The $t\bar{t}H(H \to b\bar{b})$ analysis has a low signal-to-background ratio and a large $t\bar{t}+\geq 1b$ background which is difficult to distinguish from $t\bar{t}H$. Therefore, we make use of MVA techniques in the signal regions in order to get as much information as possible from each event. This helps to separate signal from background events which in turn improves the sensitivity of the analysis. In the control regions, we do not care about separating signal from background; in these regions we use the event yield as input to the fit. The two $t\bar{t}+\geq 1c$ control regions from the semileptonic channel are, however, an exception. In order to control the $t\bar{t}+\geq 1c$ background better, the scalar sum of the p_T of all jets, H_T^{had}, is used in the fit for these regions. The resolved channels use several layers of MVA techniques whereas the boosted channel uses only one method: the BDT.

5.7.1 Boosted Decision Tree

A decision tree is a classifier structured as a binary tree in which cuts on certain input variables are applied at each node. The input variables are designed to distinguish signal from background events. At each node, the cuts are optimised to achieve the best splitting of events. The tree also chooses which variable to cut on at each node, based on which gives the best separation between signal and background for this particular node. This means that the same variable can occur at multiple nodes, and some input variables might not be used at all in the tree. A schematic of a basic decision tree is shown in figure 5.14. Events pass through a succession of nodes until either the maximum tree depth is reached, or too few events remain in a node. They are then labelled as signal (S) or background (B) when they reach the end of the tree, depending on the majority of events that end up in the final leaf nodes.

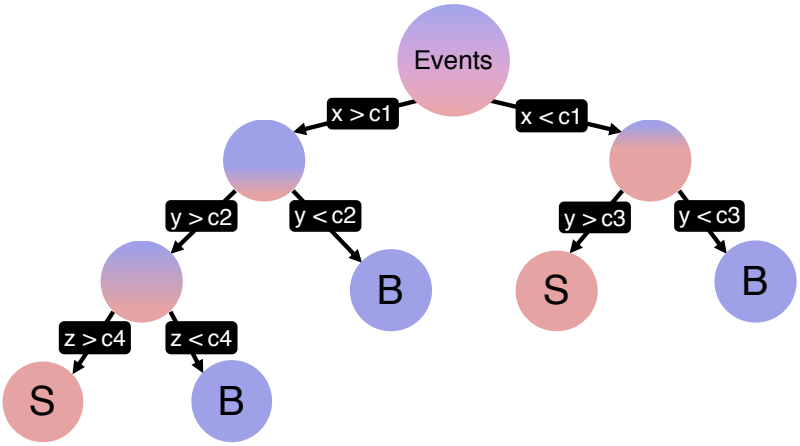


Figure 5.14: Schematic of a basic decision tree. An event passes through a series of nodes where a binary cut on a discriminating variable (x, y, z) is applied. The nodes at the very end of the tree are labelled as signal (S) or background (B) depending on the majority of events that end up in these nodes.

The cuts in each node are optimised using the *Gini Index* which is defined as $G = P(1-P)$, where P is the signal purity in a node given by the ratio of the number of signal events to the total number of events in that node [114]. This index is zero for a node completely comprised of signal or background events. Since a cut selecting background events very well is as valuable as a cut very efficient in selecting signal, the aim is to minimise the overall increase in G at each step in the tree. After a cut is applied in a node, the G values of the daughter nodes are weighted by their fraction of events and added up. This combined G index is then compared to the initial parent node's G value and minimised.

Boosting the decision tree classifier extends the concept of one tree to many trees which are used together to create a *forest*. This results in a better MVA performance and more robustness against statistical fluctuations in the training samples. Each tree in the classifier is added iteratively and uses the results of the previous tree to improve its own training. The events in the training sample are weighted at each tree according to the output of the previous one. Each tree is also assigned a weight according to its performance, and the final BDT classifier is defined as the weighted average over all trees in the forest. The output of this final classifier is a distribution in which one side contains mostly background events and the other side contains mostly signal. An optimal cut on the BDT output is defined which classifies all events below this cut as background and above this cut as signal.

We use the Adaptive Boost (AdaBoost) algorithm in which events that were misclassified in the previous tree are assigned larger weights in the training of the following tree. This assures that future trees concentrate on these events. The weights of misclassified events are multiplied by the *boost weight* α after which the weights of the entire sample are renormalised to keep a constant sum of weights. The boost weight is calculated as a function of the misclassification rate, err, according to [114]:

$$\alpha = \frac{1-\text{err}}{\text{err}}. \tag{5.5}$$

The AdaBoost algorithm performs well on an ensemble of weak classifiers, i.e. a large forest

of shallow trees. Therefore, we add an exponent β to the boost weight ($\alpha \rightarrow \alpha^{\beta}$) which is a parameter controlling the learning rate of the algorithm. Applying a $\beta < 1$ makes the algorithm learn slower which enhances the performance when combined with a large number of trees. In this way, we avoid *overtraining* the BDT.

A BDT is overtrained when it is over-optimised on the training sample and therefore cannot perform well on another, independent, data sample. This can occur when the classifier learns too many features specific to the training sample (e.g. features due to statistical fluctuations) which leads to inflexibility when it is applied to new samples. The use of a forest of trees instead of a single tree helps to mitigate this effect. Another way to minimise this problem is to use only shallow trees with a few nodes which are less likely to be overtrained. A large forest of these weakly classifying trees gives a good performance. We apply a cross-validation strategy in order to check for overtraining of our BDT. In order to execute the cross-validation we split the training sample into two samples, A and B, of equal size. We then train the BDT on one half (sample A) and test it on the other half (sample B), and vice versa. This ensures that the performance of the BDT is never extracted from the same events it was trained on. We compare the performance of both of these tests; if they are equal we are safe from overtraining. The BDTs resulting from the two different trainings are combined together in order to make use of the full statistics available. The BDT output acquired from the training on sample A is only applied to the events in sample B, and vice versa. The observed data events are also split into two groups of equal size; the BDT from training on simulated sample A is applied to the one half and the training of simulated sample B on the other half.

The BDTs in this analysis are implemented using the TMVA package [114] which outputs a variable ranking after each training session that tells us which of the variables was most important in the training. The ranking takes into account the number of times the variable was used to make cuts in the tree nodes, weighted by the separation achieved at each node and the number of events in the node. The separation of a single variable or a multivariate classifier is calculated as a sum over each of its bins according to:

$$\langle S^2 \rangle = \frac{1}{2} \sum_{\text{bins}} \frac{(n_s - n_b)^2}{n_s + n_b}, \tag{5.6}$$

where n_s and n_b are the number of signal and background events in each bin.

TMVA also provides a matrix displaying the linear correlation coefficients between each of the variables. The correlation between two variables X and Y is computed as

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{5.7}$$

where $\text{cov}(X,Y)$ is the covariance between the two variables and σ_X (σ_Y) the variance of variable X (Y). The correlation coefficients range from -1 to 1 and are symmetric in X and Y. A coefficient of 0 indicates completely independent variables and any value $\neq 0$ indicates some form of linear relationship between the variables. However, this coefficient does not tell us about higher order relationships that might exist between variables.

The performance of a BDT can be quantified by its Area Under the ROC Curve (AUC). The Receiver Operating Characteristics (ROC) curve plots the signal efficiency as a function of the background rejection. The area under this curve tells us how well our multivariate analysis is capable of distinguishing signal from background events, with a larger area indicating a better performance. A schematic of a ROC curve is shown in figure 5.15. A random classifier has a 50% chance of getting the classification right and its AUC is 0.5; a perfect classifier would have an AUC of 1. Compared to the random classifier, classifier A has a higher background rejection at each signal efficiency and has a larger area under its curve. Classifier B performs even better than A and has the largest AUC of all three.

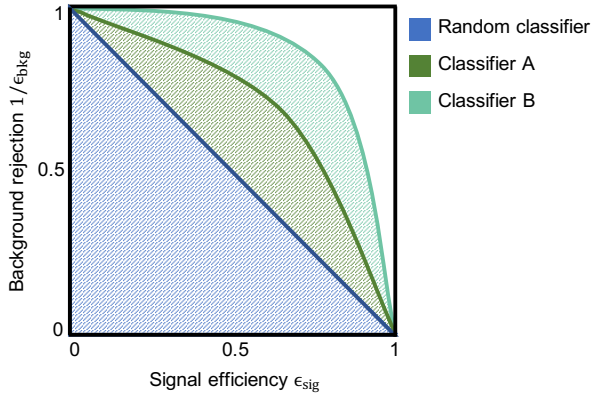


Figure 5.15: Schematic of different ROC curves that quantify the performance of a classifier such as a BDT. A random classifier has a 50% chance of getting the classification right and has an AUC of 0.5. The classifiers A and B perform better and have larger areas covered under their curves.

5.7.2 Boosted $t\bar{t}H(H \to b\bar{b})$ MVA techniques

The boosted analysis makes use of a classification BDT to separate signal from background events. The BDT is trained on a signal sample of $t\bar{t}H$ events and a background sample including $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$, $t\bar{t}+$ light, and $t\bar{t}+V$ events. All events in the training are weighted according to their full MC weight which includes the normalisation to data. The boosted filtered $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ samples described in section 5.4.2 are included in the training to ensure a good statistical coverage of the high-p_T phase-space. Since my signal region design was chosen for the analysis, I was responsible for the BDT definition and optimisation presented here.

BDT settings

A systematic study was carried out in order to understand the extent of the effect of each of the BDT settings on its performance. For this study, a loose event selection was chosen in which we require three $R = 0.4$ jets of which at least two are b-tagged at the 77% WP, and one large $R = 1.0$ trimmed jet with $m > 50$ GeV. Ten input variables were chosen that were optimised for this event selection and give a good separation between signal ($t\bar{t}H$) and background ($t\bar{t}+$ light, $t\bar{t}+\geq 1c$, $t\bar{t}+\geq 1b$, $t\bar{t}+V$) events. The BDT settings that were studied are listed in

table 5.10, along with their nominal value. Each setting is varied while the other settings are kept at their nominal value.

BDT setting	Nominal	Range	Description
AdaBoost β	0.15	[0.025, 0.25]	Learning rate for AdaBoost algorithm
Maximum tree depth	3	[1, 8]	Maximum depth of the decision tree allowed
Minimum node size	5%	[1%, 20%]	Minimum % of training events required in a leaf node
Number of trees	400	[50, 700]	Number of trees in the forest

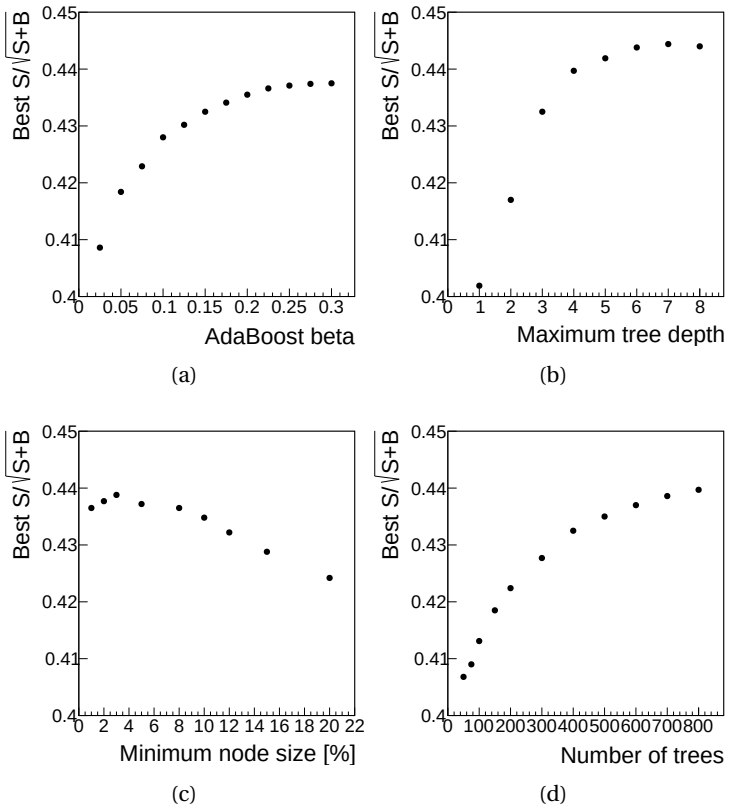
Table 5.10: The four BDT settings that were studied in order to assess their impact on the BDT performance.

Two performance metrics were chosen in order to assess the influence of each of the four settings on the BDT. Firstly, the best $S/\sqrt{S+B}$ achieved by the BDT is studied in figure 5.16. For the minimum node size (c), the optimal performance is reached around $2-3\%$, after which the $S/\sqrt{S+B}$ drops. It is expected that a large minimum node size decreases the performance since the trees would be terminated very quickly. For the other BDT settings, an increase in performance is observed with an increase in the settings. The gain of this increase flattens off at higher values, indicating that extending these studies further would not make a significant difference. The performance peaks at a maximum tree depth of 7 (b), but we need to take into account the effects of overtraining and would therefore need to pick a lower value for this.

The second metric chosen to assess the effect of the BDT settings is the separation power of the highest ranked variable; it is shown in figure 5.17. For this metric we see the opposite behaviour compared to the $S/\sqrt{S+B}$: the separation power decreases when we increase the settings. This effect is not very pronounced for the minimum node size but is very apparent for the other three settings. The two metrics together show that there is a trade-off in varying the BDT settings, where one can gain some $S/\sqrt{S+B}$ while simultaneously losing some discrimination power of the input variables. This study was repeated for the final boosted event selection and BDT variable selection chosen for the 2018 paper; similar results were found. The final settings for the boosted $t\bar{t}H(H\rightarrow b\bar{b})$ analysis BDT were optimised according to this study and chosen as shown in table 5.11.

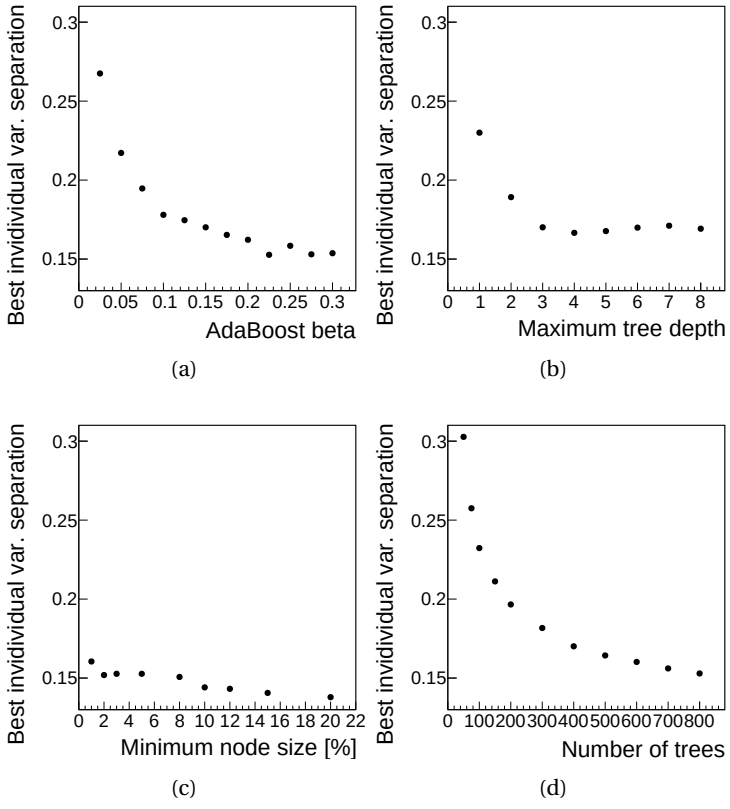
BDT setting	Optimised value
AdaBoost beta	0.3
Maximum tree depth	3
Minimum node size	2%
Number of trees	700

Table 5.11: The final BDT settings optimised for the boosted SR and BDT variable selection of the 2018 paper.



(a) (b) (c) (d)

Figure 5.16: The effect of changing four of the possible BDT settings on the best $S/\sqrt{S+B}$ that the BDT achieves.



(a) (b) (c) (d)

Figure 5.17: The effect of changing four of the possible BDT settings on the separation power achieved by the highest ranked variable in the BDT.

BDT input variables

Many variables were considered as input to the boosted BDT. New variables had to be constructed because not all JSS variables that were traditionally used in the boosted $t\bar{t}H$ analysis on trimmed jets are well-defined for the reclustered jets. As explained in section 5.5.1, any JSS variable that needs to be calculated directly from the jet energy clusters would need an extra set of uncertainties for the reclustered jets that are currently not defined. However, the jet mass and k_T splitting scales can be used on reclustered jets because their calculation can simply be made using the subjets as inputs directly instead of using the jet topocluster constituents.

A first list of input variables was constructed containing a few dozen variables that have potential to discriminate between the $t\bar{t}H$ signal events and $t\bar{t}$+ jets background events. This list includes several jet kinematics of the small and large jets in the events, substructure variables, b-tagging variables constructed from the pseudo-continuous b-tagging scale, and jet multiplicities. In order to use a variable in the analysis, it needs to be well modelled by MC simulation data, have a good discrimination power, and have a low correlation with other variables used in the BDT.

In order to narrow down the list of potential BDT variables, an elimination procedure is applied as illustrated in figure 5.18. Starting from the large list of variables, the BDT is trained and the variable ranking is examined. The lowest ranked variable is removed from the list, and the BDT is trained again. This process is repeated until we reach a list of about 20 variables. At this stage, we can start checking some of the other figures of merit for the variable performance, such as their correlation with other variables as computed by equation 5.7. We strive for a low correlation between all BDT variables in order to exploit as much information from our events as possible. If two variables have a correlation of 30% or above, both of the variables are taken out of the training one by one to check which of the two achieves a better overall performance in the BDT. The one performing less well is then discarded. With this elimination method we reach a list of 11 variables performing well inside the BDT: see table 5.12.

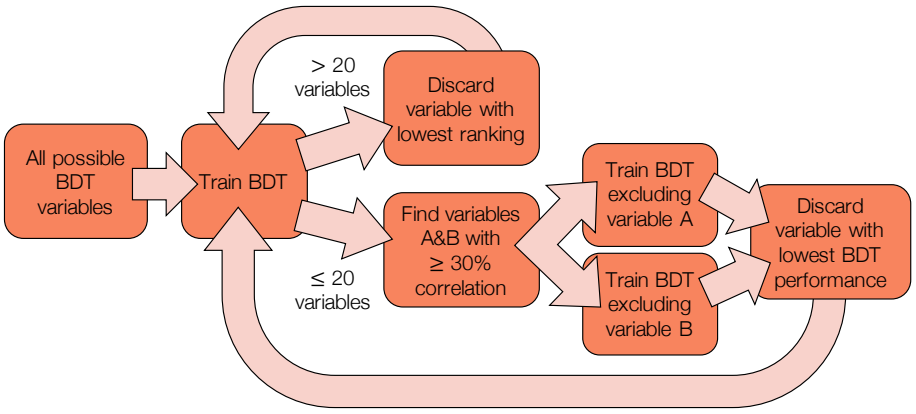


Figure 5.18: The procedure for the selection of BDT variables.

The first 10 variables in this list were constructed specifically for this analysis and signal region. The variable $\Delta R_{\text{bb max } p_T}$ has already been used before in the resolved analysis since it

Variable	Explanation
1. $w_{b\text{-tag}}$	Sum of pseudo-continuous b-tagging scores of all jets
2. $w^{\text{add}}_{b\text{-tag}}/w_{b\text{-tag}}$	Sum of pseudo-continuous b-tagging scores of all additional jets divided by $w_{b\text{-tag}}$
3. m_{Higgs}	Mass of the Higgs candidate jet
4. Higgs $\sqrt{d_{12}}$	First k_T splitting scale of the Higgs candidate jet
5. Top $\sqrt{d_{12}}$	First k_T splitting scale of the top candidate jet
6. $\Delta R_{\text{H, lep}}$	ΔR between the Higgs candidate jet and the lepton
7. $\Delta R_{\text{H, t}}$	ΔR between the Higgs candidate jet and the top candidate jet
8. $\Delta R_{\text{H, }b^{\text{add}}}$	ΔR between the Higgs candidate jet and the leading additional b-jet
9. $\Delta R_{\text{t, }b^{\text{add}}}$	ΔR between the top candidate jet and the leading additional b-jet
10. $\Delta R_{bb\text{ in H}}$	ΔR between the two leading b-tagged subjets of the Higgs candidate jet
11. $\Delta R_{bb\text{ max }p_T}$	ΔR between the two leading b-jets in the event

Table 5.12: The 11 potential input variables to the boosted BDT after the first elimination procedure. The additional jets refer to any jets that are not subjets of the Higgs and top candidate large jets.

does not require any boosted objects. The first two variables are constructed from the pseudo-continuous b-tagging scale described in section 4.3.4. For $w_{\text{b-tag}}$, all jets in the event are assigned an integer value depending on the b-tagging WP that they pass and these values are then summed up. The variable $w^{\text{add}}_{\text{b-tag}}/w_{\text{b-tag}}$ adds up just the b-tagging scores of the jets outside of the top and Higgs candidate jets, and takes the ratio with the first b-tagging variable. These b-tagging variables provide a very good separation. Variables 3 to 5 are all jet substructure variables (see section 4.4.1) and 6 to 10 are angular variables specific to the boosted objects identified in our signal region.

From the 11 candidate variables, a further selection is made by looking at two figures of merit for the BDT performance: the maximum $S/\sqrt{S+B}$ achieved, and the AUC. The final selection of BDT input variables is based on the study summarised in table 5.13. Six different sets of variables were picked from the 11 listed in table 5.12 and their BDT performance compared. The sets of variables are based on the highest ranking variables with the smallest correlation between them. Any sets of less than eight variables are not considered because they lose significant sensitivity compared to the others. The first row shows the results when using all of the 11 variables; it has the best $S/\sqrt{S+B}$ and one of the best AUC scores, but the performance differences between the six sets are marginal. It is beneficial to use as few input variables as possible since the modelling of each of them needs to be checked carefully and using many variables in a low statistics sample can lead to overtraining. The final option using eight variables was chosen because it uses the least variables while still achieving a similar performance to the sets with more variables.

The importance ranking of the final eight BDT input variables is shown in table 5.14. The distributions of these variables are shown in figures 5.19–5.22 alongside the separation they

Input variables	Number of variables	AUC	Max $S/\sqrt{S+B}$
1-11	11	0.739	0.827
1-3, 5-11	10	0.736	0.813
1-9, 11	10	0.741	0.820
1-3, 5-9, 11	9	0.737	0.809
1-3, 5-10	9	0.737	0.816
1-3, 5-9	8	0.737	0.811

Table 5.13: The performance of the BDT measured by the AUC and maximum achieved $S/\sqrt{S+B}$ for six different sets of variables picked from table 5.12.

Ranking	Variable
1	$w_{b\text{-tag}}$
2	m_{Higgs}
3	$\Delta R_{\text{H, t}}$
4	$\Delta R_{\text{H, }b^{\text{add}}}$
5	$\Delta R_{\text{H, lep}}$
6	$\Delta R_{\text{t, }b^{\text{add}}}$
7	$w^{\text{add}}_{b\text{-tag}} / w_{b\text{-tag}}$
8	Top $\sqrt{d_{12}}$

Table 5.14: The ranking of the eight input variables to the boosted BDT.

achieve between signal and background events. All variables show good agreement between data and MC simulation within the uncertainty bands. The uncertainty on data is statistical only and on MC it includes statistics and all systematics; these will be described in chapter 7. As expected, we see a peak around the Higgs mass in figure 5.20(a), and a peak around half the top mass in 5.20(c) (see section 4.4.1 for an explanation of the first splitting scale JSS variable). Figure 5.21(c) tells us that the Higgs and top jets are most often produced back-to-back, which is expected since they are usually the only two boosted objects in the events. The largest individual variable separation of 16.2% is achieved by the $w_{\text{b-tag}}$ variable in figure 5.19(b), after which comes the Higgs mass in figure 5.20(b) with a separation of 6.94%.

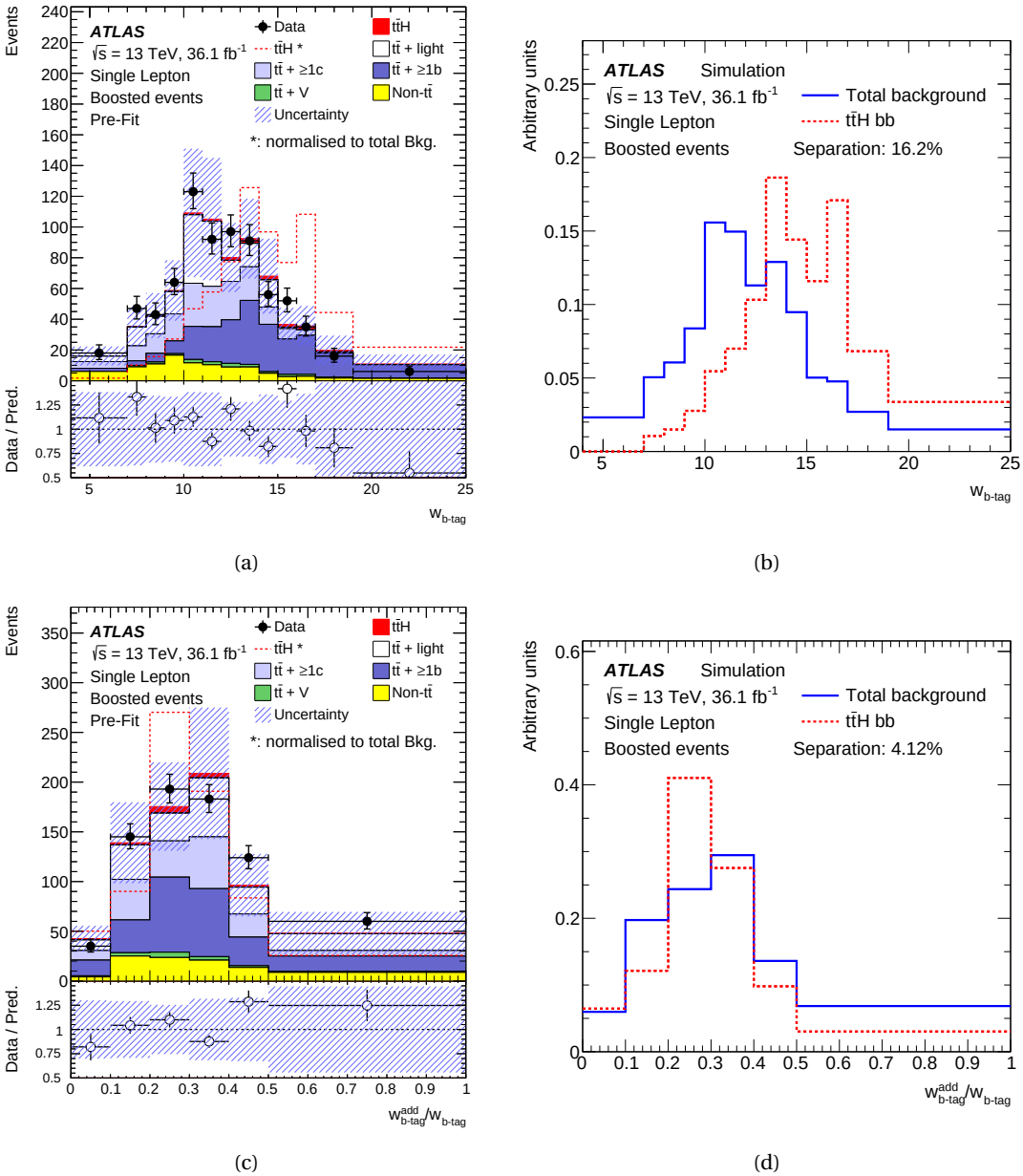
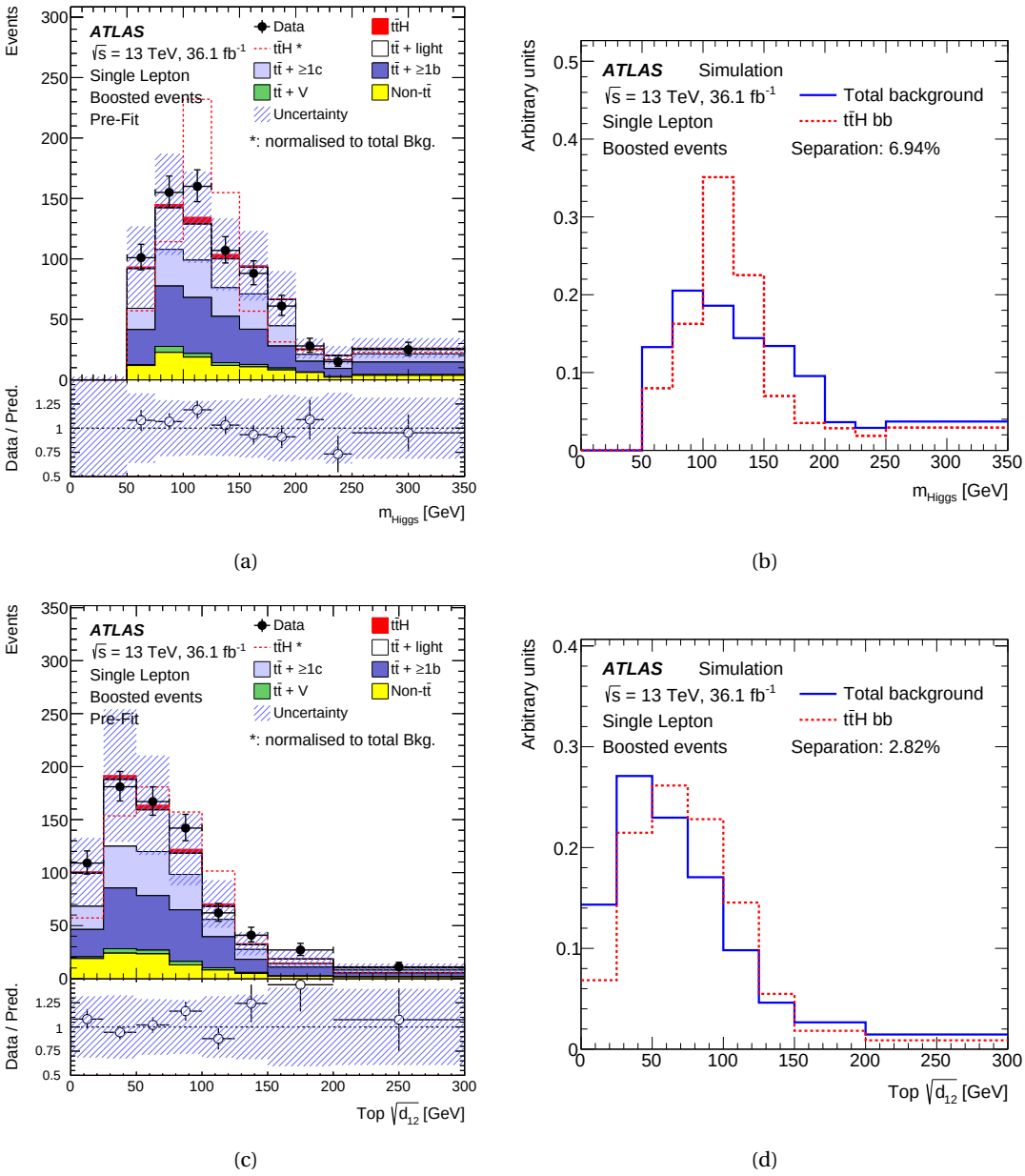


Figure 5.19: Two of the eight variables used in the boosted classification BDT. The left column shows the distribution where the $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section. The signal is also shown in a red dashed line where it is normalised to the total background prediction. The dashed blue bars indicate the total uncertainty including systematics which will be explained in chapter 7. The right column shows the separation between signal and background achieved by each variable.



(a) (b)

(c) (d)

Figure 5.20: Two of the eight variables used in the boosted classification BDT. The left column shows the distribution where the $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section. The signal is also shown in a red dashed line where it is normalised to the total background prediction. The dashed blue bars indicate the total uncertainty including systematics which will be explained in chapter 7. The right column shows the separation between signal and background achieved by each variable.

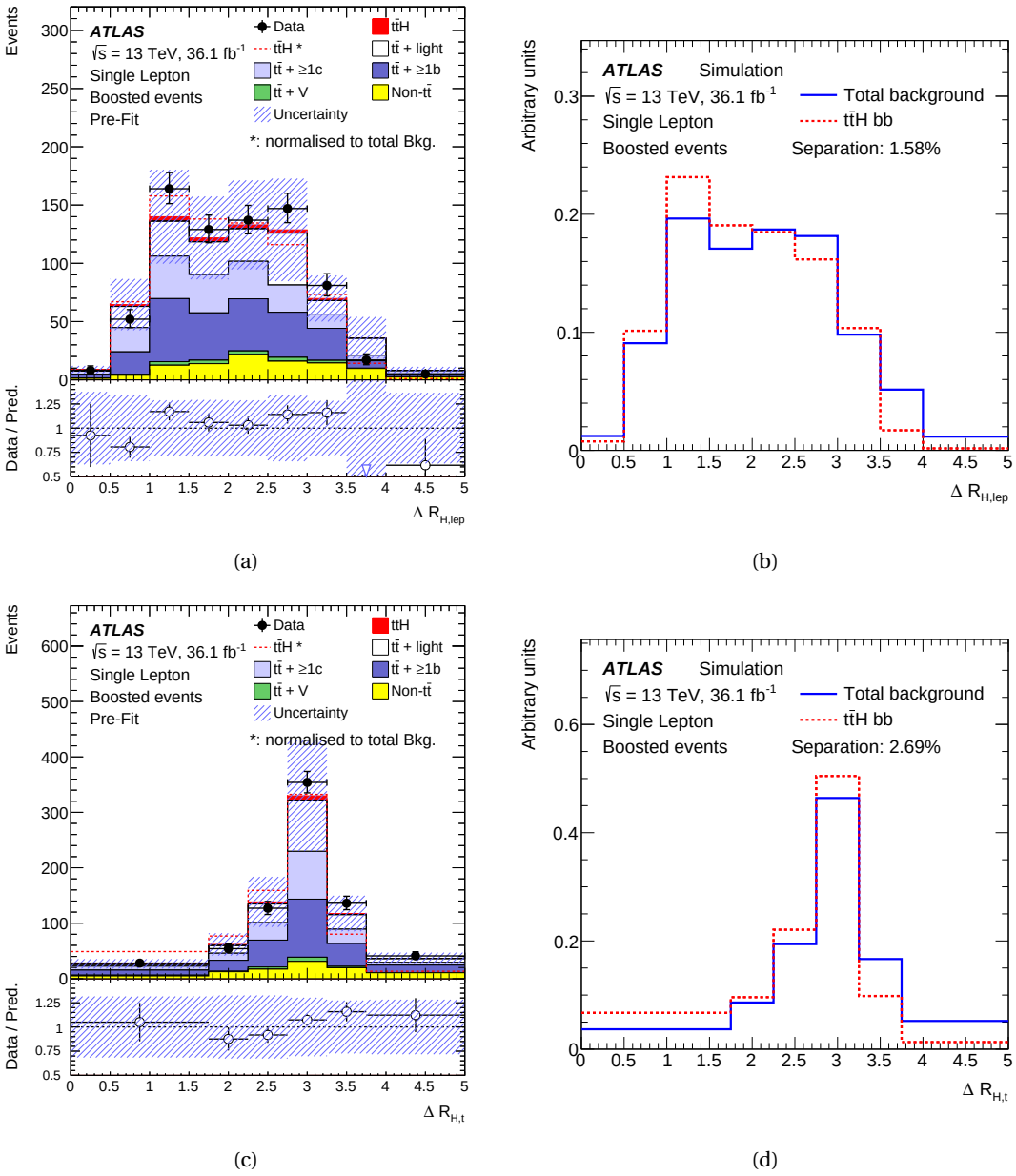
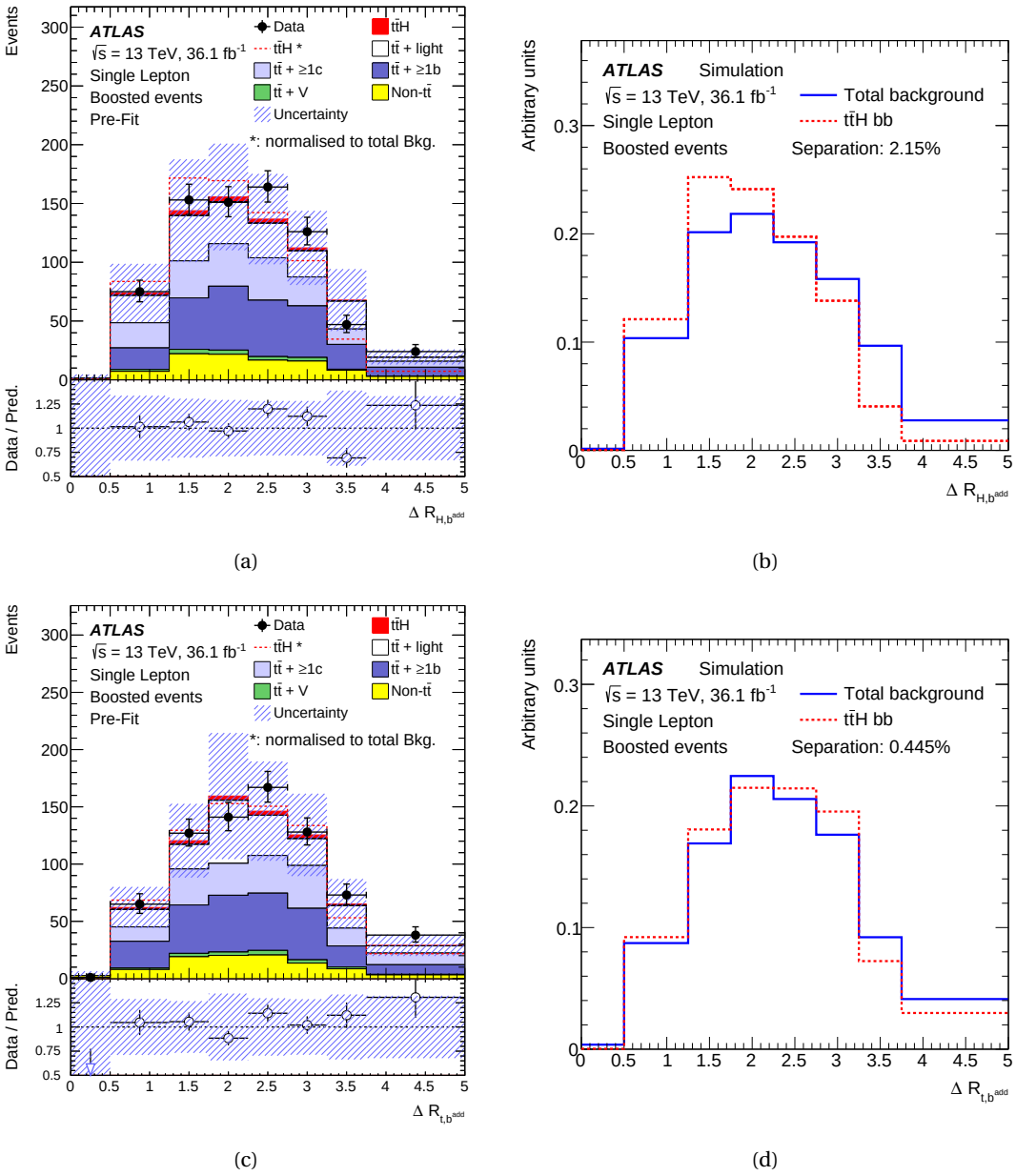


Figure 5.21: Two of the eight variables used in the boosted classification BDT. The left column shows the distribution where the $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section. The signal is also shown in a red dashed line where it is normalised to the total background prediction. The dashed blue bars indicate the total uncertainty including systematics which will be explained in chapter 7. The right column shows the separation between signal and background achieved by each variable.

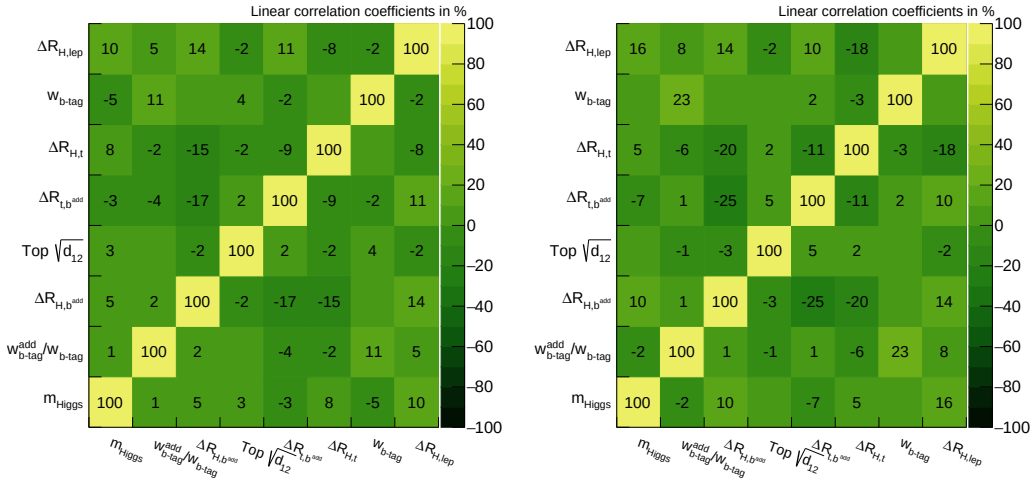


(a) (b)

(c) (d)

Figure 5.22: Two of the eight variables used in the boosted classification BDT. The left column shows the distribution where the $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section. The signal is also shown in a red dashed line where it is normalised to the total background prediction. The dashed blue bars indicate the total uncertainty including systematics which will be explained in chapter 7. The right column shows the separation between signal and background achieved by each variable.

The linear correlations between each of the eight variables is shown in figure 5.23 for background events (a) and signal events (b). As mentioned before, we strive for a low correlation between all BDT variables in order to exploit as much information from our events as possible. The strongest correlation we have in our multivariate analysis is of −25% between the $\Delta R_{\mathrm{t},b^{\mathrm{add}}}$ and $\Delta R_{\mathrm{H},b^{\mathrm{add}}}$ variables in signal events.



(a) Correlations in background events

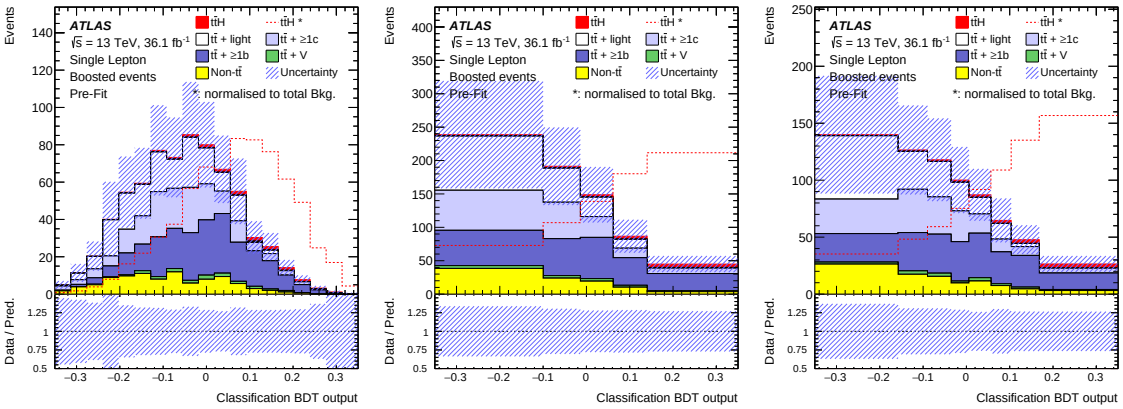
(b) Correlations in signal events

Figure 5.23: Linear correlations between the eight variables chosen for the boosted BDT.

BDT binning

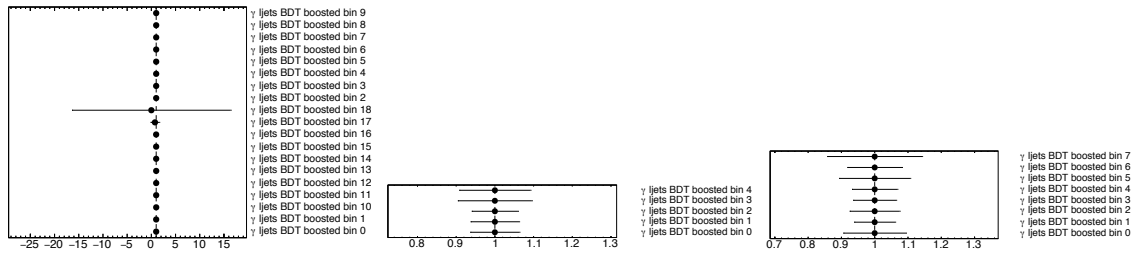
After the settings and input variables of the classification BDT are fixed, the last important step is to choose an appropriate binning of its output distribution. We aim to have a statistical uncertainty below 20% in each BDT bin whilst also keeping the BDT performance as high as possible. Originally, a binning of 19 equal bins was chosen (from here on marked as “default”), but this results in very low statistics in both the lowest and highest bins. An auto-binning algorithm was tested which automatically optimises the BDT performance given the number of bins to use on the left- and right-hand side of the optimal BDT cut. The algorithm tested is TransfoD with two different bin multiplicities. Firstly, the option "3,2" is used which means the algorithm constructs 3 bins on the left-hand side of the optimal BDT output cut, and 2 bins on the right-hand side of this cut. The second TransfoD binning is taken to be "4,4" which constructs 4 bins on the background-like side of the BDT and 4 bins on the signal-like side.

In order to study the impact of the BDT binning on its bin uncertainties and overall performance, a fit to Asimov data (see section 6.3) was performed over the boosted signal region only, with all systematics included. The pre-fit BDT output distribution for the three different binning options is shown in figure 5.24. The statistical uncertainty in each bin is shown in figure 5.25 for the three different binning options. The default binning in subfigure (a) has an extremely large uncertainty in the last (18th) BDT bin. For both the TransfoD32 and TransfoD44 binning options the statistical uncertainties remain well below the desired 20%.



(a) Default (b) TransfoD32 (c) TransfoD44

Figure 5.24: BDT output distributions with three different binning choices, shown pre-fit.



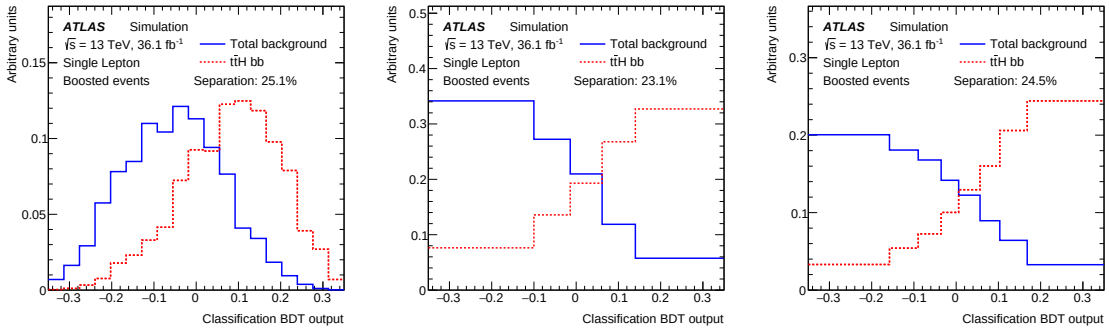
(a) Default (b) TransfoD32 (c) TransfoD44

Figure 5.25: The statistical uncertainties associated with each BDT bin for three different binning choices. The horizontal bars indicate the statistical uncertainty in the BDT bin, where a length of 0.1 indicates an uncertainty of 10%.

The separation power of the BDT remains very similar between the three different binning options, as shown in figure 5.26. TransfoD44 does result in a slightly better performance than TransfoD32. The total uncertainties on the signal strength and normalisation of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds are shown in table 5.15. We see that the TransfoD44 has smaller uncertainties on these three parameters than TransfoD32. Overall, the TransfoD44 was chosen as the best BDT binning option because it gives a good balance between low bin uncertainties and high BDT performance.

Binning	$\mu_{t\bar{t}H}$	**k**$(t\bar{t}+\geq 1b)$	**k**$(t\bar{t}+\geq 1c)$
Default	$+2.1, -1.8$	$+0.7, -0.4$	$+1.1, -0.8$
TransfoD32	$+3.4, -3.7$	$+1.0, -0.6$	$+1.4, -1.0$
TransfoD44	$+2.7, -3.0$	$+0.9, -0.5$	$+1.3, -1.0$

Table 5.15: The total uncertainties on the signal strength and the two normalisation factors k, shown for the three different BDT binning choices.



(a) Default (b) TransfoD32 (c) TransfoD44

Figure 5.26: The BDT separation performance for three different binning choices.

Negative weights in the BDT training

The MC samples used in this analysis all have a fraction of events with negative weights. The number of events with negative weights is shown in table 5.16. The fractions are especially large for the $t\bar{t}H$ signal and $t\bar{t}+V$ backgrounds since they both have their matrix element generated by MG5_aMC@NLO. The TMVA [114] package does not allow for negative weights in the training of the BDT and therefore a method must be established on how to deal with these events. We have chosen to use the absolute value of the event weights in the BDT training because this does not significantly change the distributions of the input variables. A comparison between the usage of the absolute value and the nominal event weight is shown in figure 5.27 for $t\bar{t}H$ and $t\bar{t}$ events. Three of the boosted BDT variables are shown; the results are very similar for the other 5 variables. The $t\bar{t}H$ sample shows larger fluctuations because it has lower statistics than all the background samples combined and a larger percentage of negative weights.

Sample	Total number of events	Number of events with negative weight	Percentage of events with negative weight
$t\bar{t}H$	15661	4972	32
$t\bar{t}+\geq 1b$	546	3	0.5
$t\bar{t}+\geq 1c$	133	1	0.8
$t\bar{t}+$ light	249	1	0.4
$t\bar{t}+\geq 1b$ boosted	19888	98	0.5
$t\bar{t}+\geq 1c$ boosted	2338	10	0.4
$t\bar{t}W$	1265	248	20
$t\bar{t}Z$	2747	887	32

Table 5.16: The number of negative weights found in the signal $t\bar{t}H$ sample and each of the $t\bar{t}$ background samples used in the training of the BDT.

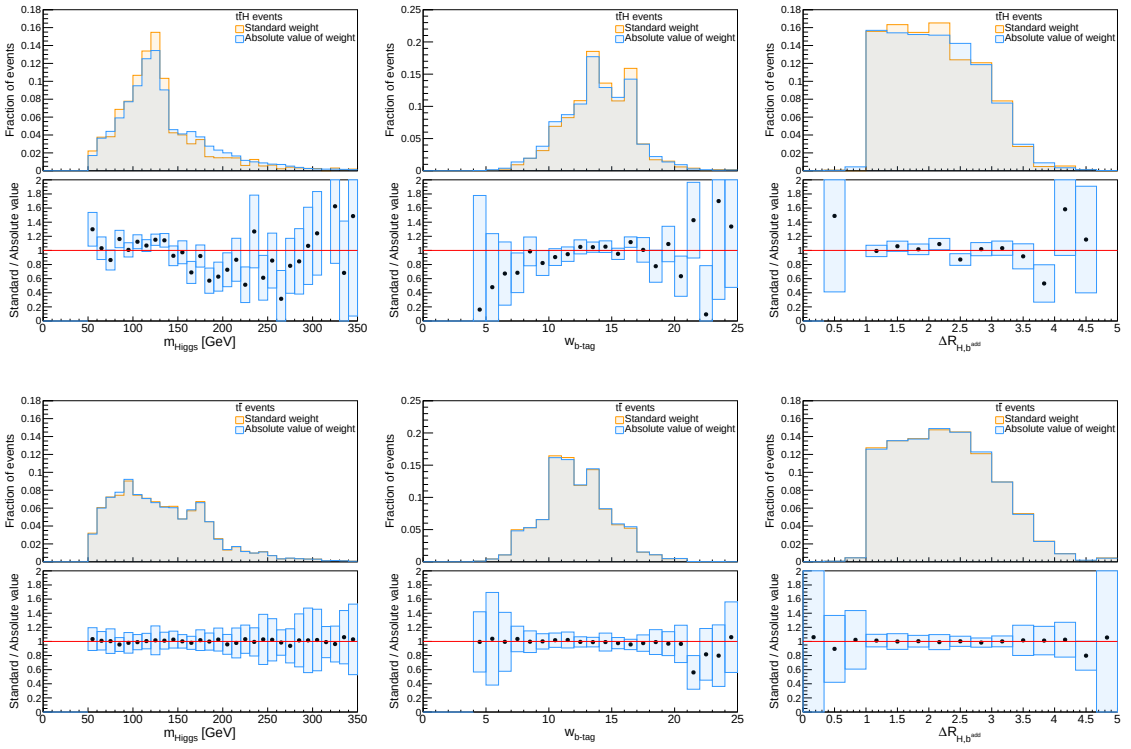
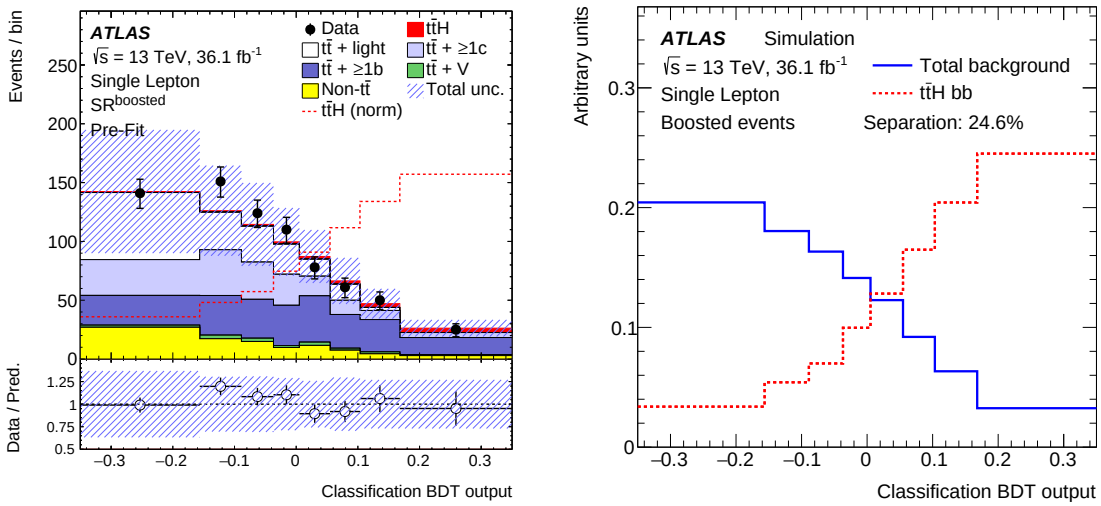


Figure 5.27: Comparison between using the nominal weights for all events and using the absolute value of the event weights for three of the boosted classification BDT variables. The top row shows the comparison for $t\bar{t}H$ signal events and the bottom row for all $t\bar{t}$ background events as shown in table 5.16 combined.

BDT performance

The final BDT designed for the boosted signal region is shown in figure 5.28 along with its separation power between signal and background events. A good agreement between data and MC is observed and the separation power achieved is 24.6%. While the separation power gives an indication of the BDT performance, it is heavily dependent on the chosen binning. Therefore, the performance of the BDT is best quantified by the area under its ROC curve.

As explained in section 5.7.1, we apply a cross-validation strategy in order to check for overtraining of the BDT. The training sample is split in two equal parts by separating the odd numbered events from the even numbered ones. We then train the BDT on the events in the "even" group and test them on the "odd" group, and vice versa. The BDTs resulting from both trainings are combined in the final classifier in order to make use of the full available statistics. The BDT output acquired from the training on the "even" sample is only applied to the "odd" events, and vice versa. The ROC curves for both the odd-on-even and even-on-odd training/testing cases are shown in figure 5.29. The shapes and AUCs of both curves are compatible, indicating no overtraining or bias in the multivariate analysis.



(a) (b)

Figure 5.28: Classification BDT output from the semileptonic boosted signal region (a) and its separation power between signal and background events (b). In (a), the $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section. The signal is shown also in a red dashed line where it is normalised to the total background prediction.

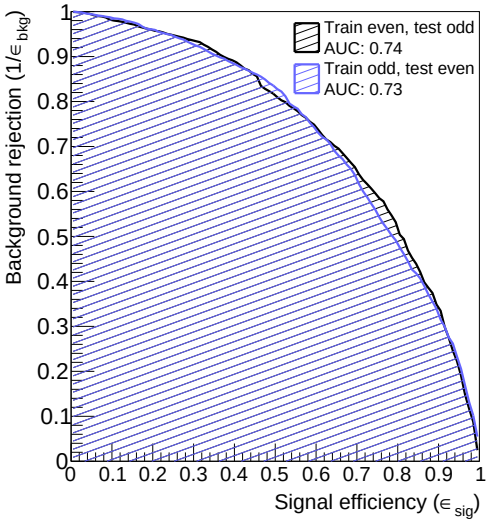


Figure 5.29: The ROC curves for the boosted BDT for the two cases of training on even events and testing on odd events, and vice versa.

5.7.3 Resolved $t\bar{t}H(H \to b\bar{b})$ MVA techniques

The $t\bar{t}H(H \to b\bar{b})$ final state has many jets coming from the decay of the Higgs boson and the two top quarks, as well as from additional radiation. The boosted channel has simplified combinatorics because some of these jets are captured together into large jets which we can then identify as the Higgs boson or the hadronic top quark. The resolved channel, however, does not have this advantage and therefore uses multivariate techniques to deal with the many possible combinations of jets in the final state to reconstruct the Higgs boson and top quark candidates. These techniques are collectively referred to as the reconstruction stage, which is the first step in the two-stage MVA strategy implemented in the resolved analysis.

The first MVA stage attempts to reconstruct the final state of the signal and background processes from the jets, leptons, and MET in the events. This stage employs three different MVA techniques: the reconstruction BDT, the likelihood discriminant (LHD), and the Matrix Element Method (MEM). The second stage builds a classification BDT similar to the boosted analysis, using input variables constructed in the first stage as well as kinematic variables and b-tagging variables. Each signal region uses a classification BDT as its final discriminant. The resolved techniques are briefly summarised here.

Reconstruction BDT

The reconstruction BDT is used to match jets in the event to final state partons, thereby building the Higgs and top candidates. It is trained on simulated $t\bar{t}H$ events to distinguish between correct and incorrect jet assignments using masses, angular separations of objects, and kinematic variables. This reconstruction BDT is employed in all the resolved single-lepton and dilepton signal regions. It uses the properties of the reconstructed Higgs and top candidates to define discriminating variables for $t\bar{t}H$ against $t\bar{t}$+ jets which are used as inputs to the classification BDT. This method correctly reconstructs the Higgs boson in 48% of selected $t\bar{t}H$ events in the semileptonic $\mathrm{SR}_1^{\geq 6j}$ region and in 49% for the dilepton $\mathrm{SR}_1^{\geq 4j}$ region. This is comparable to the boosted region which tags the correct Higgs candidate in 47% of selected $t\bar{t}H$ events.

Likelihood discriminant

The likelihood discriminant is computed analogously to reference [115]. It gives the probability for signal ($t\bar{t}H$) and background ($t\bar{t}+ \geq 1b$) hypotheses using 1D distributions of discriminating variables such as the invariant mass and angular separations. The LHD is defined as

$$\mathrm{LHD} = \frac{p_{\mathrm{sig}}}{p_{\mathrm{sig}} + p_{\mathrm{bkg}}}, \tag{5.8}$$

where p_{sig} (p_{bkg}) is the probability density function of a given event under the signal (background) hypothesis. The LHD is employed in all resolved single-lepton signal regions and results in one variable to be used as input to the classification BDT.

Matrix Element Method

The MEM is very computationally intensive and is therefore only used in the purest single-lepton signal region, $SR_1^{\geq 6j}$. It is based on the method used in the $t\bar{t}H(H \to b\bar{b})$ analysis at $\sqrt{s} = 8$ TeV [25]. The MEM discriminant is similar to the LHD, however the hypothesis testing is performed at parton level using a transfer function instead of using the reconstructed objects directly. The transfer functions map the detector quantities to parton level quantities. The MEM expresses the degree to which each event is consistent with the signal ($t\bar{t}H$) and background ($t\bar{t} + b\bar{b}$) hypotheses as likelihoods L_S and L_B. These likelihoods are calculated using matrix element calculations at parton level. This method results in a single variable to be used in the classification BDT, defined as:

$$\text{MEM}_{D1} = \log_{10}(L_S) - \log_{10}(L_B). \tag{5.9}$$

Classification BDT

The final stage in the resolved MVA strategy is the training of a classification BDT to separate $t\bar{t}H$ signal from $t\bar{t}$ background events, just as is done in the boosted channel. The TMVA [114] package is used to train both the classification and reconstruction BDTs. Inputs to the classification BDT are all the variables constructed in the first stage of the MVA strategy, plus kinematic variables and variables from b-tagging. Only variables that are well modelled by MC simulation data are used in the BDT. The outputs of the three MVA methods in the reconstruction stage described above are the most powerful in their separation in the classification BDT. The dilepton signal regions use a total of 20 input variables (including 7 variables from the reconstruction BDT) and the single-lepton signal regions use 23 variables (including the LHD, MEM discriminant, and 7 variables from the reconstruction BDT). The input variables are selected to maximise the performance of the classification BDT and not every variable is used in each region.

6 STATISTICAL ANALYSIS

In order to observe the $t\bar{t}H$ signal or set limits on this process, we perform a frequentist statistical analysis comparing our data to the SM expectation. The statistical methods used are the standard in the ATLAS and CMS experiments for searches for unseen processes, as described in references [116, 117]. The methods used in the $t\bar{t}H$ analysis are discussed in this chapter.

6.1 Hypotheses and the test statistic

Since the $t\bar{t}H$ signal has not been observed yet, we test two hypotheses in our analysis: H_0 and H_1. The former refers to the background-only hypothesis in which there is no signal present and is therefore also referred to as the null hypothesis. The latter is the hypothesis of the background-plus-signal model which is the SM prediction multiplied by a signal strength parameter μ. This signal strength is the parameter of interest in this analysis and is defined as

$$\mu = \frac{\sigma_{\text{hypothesis}}}{\sigma_{\text{SM}}}, \tag{6.1}$$

where σ is the cross-section of the process under consideration and $\mu = 1$ corresponds to the SM expectation.

Claiming an observation requires a rejection of the null hypothesis, H_0, in favour of the signal-plus-background hypothesis, H_1. The p-value is computed for the purpose of quantifying the level of agreement between our measured data and the two hypotheses. It is defined as the probability to observe data of equal or greater incompatibility with the hypothesis H than the observed level, under the assumption that H is true. This p-value is described in terms of a test statistic, q, under which the hypothesis H follows a predicted distribution: $f(q|H)$. It is given by

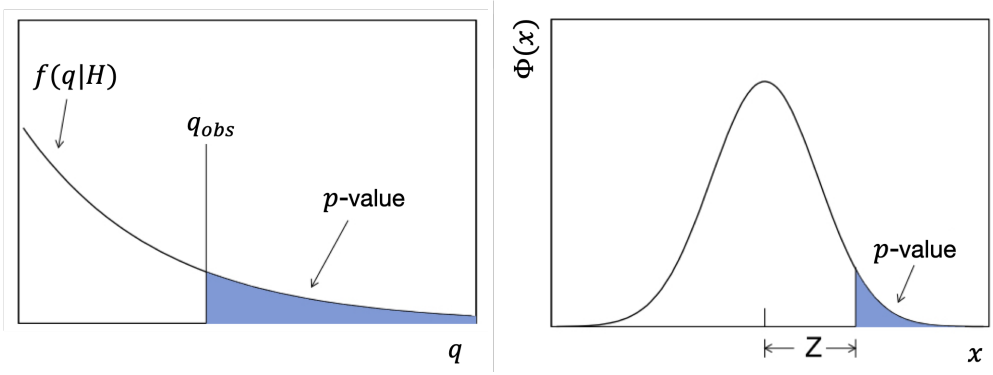
$$p\text{-value} = \int_{q_{\text{obs}}}^{\infty} f(q|H)\,dq, \tag{6.2}$$

where q_{obs} is the value of the test statistic as measured in data. The hypothesis H is excluded if its p-value is observed below a specific threshold.

Instead of quoting the p-value directly, we usually quote the significance, Z, which is defined as

$$Z = \Phi^{-1}(1-p), \tag{6.3}$$

where Φ is the cumulative standard Gaussian distribution and Φ^{-1} is its inverse. Figure 6.1 shows a schematic of the p-value defined from q_{obs} (a) and its relation to the significance (b). In particle physics, a significance of 5σ, corresponding to a p-value of 2.87×10^{-7}, is required in order to reject the background-only hypothesis and claim a *discovery*. If we reach $Z = 3\sigma$, the community speaks of *evidence* for the particular process. The rejection of an alternate hypothesis (i.e. H_1) requires a 95% confidence level which corresponds to $Z = 1.64\sigma$ and a p-value of 0.05.



(a) (b)

Figure 6.1: Schematics of the relation between the p-value and (a) the observed value of the test statistic q, and (b) the significance Z, adapted from reference [116].

6.2 Profile likelihood technique

The likelihood is defined as the probability to obtain the observed data under a given hypothesis. It is a function of the parameter of interest (μ) as well as a set of nuisance parameters (NPs) ($\boldsymbol{\theta}$) describing the systematic uncertainties. These systematics include both theoretical and experimental uncertainties. The NPs are not known a-priori but need to be fitted from the data. They provide extra degrees of freedom for the fit which it uses to correct the predicted template and match the data. This additional flexibility is needed but results in a loss of sensitivity of the analysis. The full template including the NPs and parameter of interest should be sufficiently flexible such that, for some value of all the parameters, it represents the true model.

If we want to confirm a hypothesis H, we aim to maximise the likelihood function $L(H)$. In particle physics, we use a likelihood ratio of the signal and null hypotheses as a test statistic in order to establish discovery or exclusion of an unseen process. This ratio is the most powerful test to reject a null hypothesis in favour of an alternate hypothesis at a given confidence level, as stated by the Neyman-Pearson lemma [118].

The data we use in our analysis is binned in histograms and the content of each bin is expected to follow the Poisson probability distribution. The likelihood L is therefore defined as the product of Poisson probabilities across all bins, N, of a given binned dataset:

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}))^{n_i}}{n_i!} e^{-(\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}))} \prod_{\theta_j \in \boldsymbol{\theta}} f(\theta_j), \tag{6.4}$$

where $s_i(\boldsymbol{\theta})$ and $b_i(\boldsymbol{\theta})$ are the predicted number of signal and background events in the i-th bin and n_i is the number of observed data events in that bin. The functions $f(\theta_j)$ are the penalty terms given by the probability density functions of each of the NPs. These penalty terms are usually implemented as Gaussian or Poisson priors. The NPs relevant for the analysis in this thesis will be discussed in section 7.3. We use the profile likelihood ratio to test a hypothesised value of μ:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}. \tag{6.5}$$

In the numerator, $\hat{\hat{\boldsymbol{\theta}}}$ denotes the value of $\boldsymbol{\theta}$ that maximises L for the specified μ (i.e. the *conditional* maximum-likelihood estimator). It is said that θ is *profiled*. The parameters with a single hat in the denominator signify the value of these parameters which maximise the overall likelihood (i.e. the *unconditional* maximum-likelihood estimators). The ratio assumes values between 0 and 1 (at $\mu = \hat{\mu}$), with a value close to 1 indicating a good agreement between the observed data and the hypothesised value of μ. The NPs broaden the profile likelihood which reflects the loss of information about μ due the systematic uncertainties.

The presence of our signal process, $t\bar{t}H$, can only increase the event rate on top of the background-only expected rate, which means that $\mu \geq 0$ always. We therefore define an alternative test statistic $\tilde{\lambda}(\mu)$ as

$$\tilde{\lambda}(\mu) = \begin{cases} \dfrac{L(\mu, \hat{\hat{\theta}}_\mu)}{L(\hat{\mu}, \hat{\theta})}, & \text{for} \quad 0 \leq \hat{\mu} \leq \mu \\ \dfrac{L(\mu, \hat{\hat{\theta}}_\mu)}{L(0, \hat{\hat{\theta}}_0)}, & \text{for} \quad \hat{\mu} < 0. \end{cases} \tag{6.6}$$

The test statistic for $\hat{\mu} < 0$ is defined as such since, if we find data where $\hat{\mu} < 0$, the best level of agreement between this data and any physical value of μ is when $\mu = 0$ (since $\mu \geq 0$ always). This prevents downward fluctuations from serving as evidence against the background.

For convenience, we define our test statistic as

$$q_\mu = -2 \ln \tilde{\lambda}(\mu). \tag{6.7}$$

The logarithm is added since it transforms the product in the likelihood definition to a sum which is easier to work with. The negative sign is added because the algorithms used for these calculations are better designed to find minima rather than maxima. In this form, the higher the value of q_μ, the higher the level of incompatibility between the data and the hypothesis.

6.3 Expected significance

In order to quantify the sensitivity of an experiment before performing a measurement on data, we calculate the expected significance of the experiment under assumption of the signal-plus-background hypothesis. This expected significance refers to the expected *median* significance to reject different values of μ. In the case of discovery we obtain the median significance, under the assumption of H_1 (i.e. $\mu = 1$), with which the H_0 hypothesis is rejected (i.e. $\mu = 0$).

Figure 6.2 illustrates the sensitivity of an experiment with two different hypotheses: μ (e.g. background-only hypothesis) and μ' (e.g. signal-plus-background hypothesis). The sensitivity is defined as the p-value corresponding to the median q_μ under assumption of μ'. The test statistic q_μ is given by equation 6.7 with μ set to zero in the case of discovery.

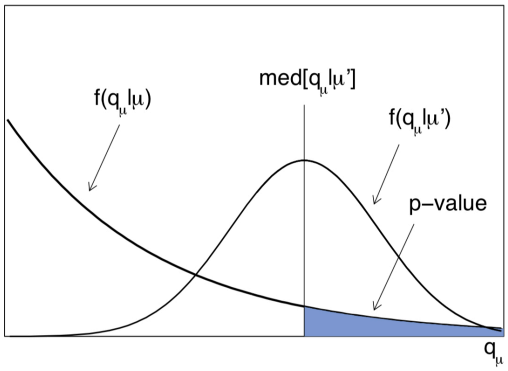


Figure 6.2: Illustration of the sensitivity of an experiment defined as the p-value corresponding to the median of q_μ under assumption of μ' [116].

The mean significance can be calculated by generating a large number of datasets based on both H_0 and H_1 and repeatedly running the analysis in order to find the mean. However, such a method would take very long and would be very computationally intensive. We therefore replace the large number of datasets with one dataset representing the 'typical' experiment: the *Asimov* dataset [116]. This dataset is defined as the dataset which gives the true values of the estimators for all parameters and delivers the desired median sensitivity. In this dataset, all statistical fluctuations are suppressed and the number of 'observed' events in each bin is set equal to the predicted number of events in that bin. The *Asimov* dataset can be constructed separately for H_0 (where $n_i = b_i$) and for H_1 (where $n_i = s_i + b_i$). These two datasets are then used to place an upper limit on the H_1 hypothesis and to get the expected discovery significance, respectively. For the expected discovery significance, we use the test statistic in equation 6.7 and set $\mu = 0$ since rejecting the null hypothesis effectively means discovering a signal. We then compute the median discovery significance assuming a strength parameter μ': $\text{med}[Z_0|\mu'] = \sqrt{q_{0,\text{A}}}$ [116]. For the upper limit, we are interested in the median exclusion significance assuming a strength parameter $\mu' = 0$: $\text{med}[Z_\mu|0] = \sqrt{\tilde{q}_{\mu,\text{A}}}$, where the test statistic is given by equation 6.8 [116].

6.4 CL_s method for upper limits

If an experiment is unable to reject the null hypothesis it can still set upper limits on the alternate hypothesis and thereby exclude regions of phase-space. Setting upper limits requires a different approach to the one taken when claiming a discovery. The expected exclusion limits are obtained from the median significance, under the assumption of H_0, with which the H_1 hypothesis is rejected. This means looking for the value of μ' where the $\text{med}[q_0|\mu']$ gives a p-value of 0.05.

In order to set the upper limits we use an alternate test statistic [116] defined as

$$\tilde{q}_\mu = \begin{cases} -2\ln\left(\tilde{\lambda}(\mu)\right), & \text{for} \quad \hat{\mu} \leq \mu \\ 0, & \text{for} \quad \hat{\mu} > \mu. \end{cases} \tag{6.8}$$

The test statistic for $\hat{\mu} > \mu$ is set to zero because an upward fluctuation of the signal does not serve as evidence against the signal. The $\hat{\mu} > \mu$ region is therefore excluded from the test's rejection region.

We set an upper limit on our parameter of interest by using the CL_s method [119, 120]. This method is used to identify the values of μ which are excluded at a confidence level of 95%. This means that, in an ensemble of experiments, 95% of the obtained confidence intervals $[0, \mu_{\text{upper}}]$ will contain the true value of μ. The method uses a modified p-value instead of directly using the p-value from the alternate hypothesis:

$$p'_{s+b} = \frac{p_{s+b}}{1 - p_b}, \tag{6.9}$$

where p_{s+b} (p_b) is the probability to obtain a result which is equal or less compatible with the signal-plus-background (background-only) hypothesis than the observed result. The CL_s method penalises the p-value of the signal-plus-background hypothesis based on the background-only probability and thus avoids false exclusion of the signal-plus-background hypothesis in cases where the analysis has no sensitivity. It takes into account the fact that the background-only and signal-plus-background distributions will overlap significantly when the expected signal yield is low. This makes the two distributions hard to distinguish from each other. In this case, a small downward fluctuation in the background might lead to a very small p_{s+b} which would lead us to reject the signal-plus-background hypothesis. However, it is the low sensitivity of the experiment ($s \ll s+b$) which leads to a false exclusion. The modified p-value is therefore used as a more conservative approach to exclude these cases. The 95% confidence level upper limit on μ corresponds to the value of μ for which $p'_{s+b} = 0.05$.

6.5 Nuisance parameters

The NPs in the fit describe the systematic uncertainties and provide extra flexibility that allows the fit to correct disagreements between the observed and expected data. Each NP, θ, has an expected value and the fit pulls it from its central value when finding the maximum-likelihood estimator of θ. To quantify how far the NP value moves away from its expectation value we define the *pull* of a NP as

$$\text{pull}(\theta) = \frac{\hat{\theta} - \theta_0}{\sigma_\theta}, \tag{6.10}$$

where θ_0 is the expected value of θ and σ_θ its standard deviation. The pulls in the fit are preferably as low as possible and should not be larger than one. Very large pulls indicate that the fit is either missing some information (e.g. some NPs are not considered) or the assumed information is partially incorrect (e.g. wrong expectation values of the NPs).

Given a sufficiently large dataset, the NPs can be constrained from the data with limited or no prior knowledge about their shape or size. If the NPs are not constrainable from the data, an auxiliary measurement or MC studies can be used to define the expected value and error of each NP. This can be performed in control regions of the experiment, or in a completely separate experiment. Even when the NPs can be constrained from the data, external knowledge which constrains the value of a NP further can be incorporated into the model.

The impact of a NP measures how much the signal strength changes as the NP is varied. The impact is defined as

$$\text{impact}(\theta) = \Delta\mu^{\pm} = \hat{\hat{\mu}}_{\theta_0 \pm \sigma_\theta} - \hat{\mu}, \tag{6.11}$$

where $\hat{\hat{\mu}}_{\theta_0 \pm \sigma_\theta}$ is the conditional maximum-likelihood estimator of μ with θ set to its expectation value plus or minus one standard deviation, and all other NPs profiled. All NPs with low impact ($< 1\%$) are discarded in the fit. We call this procedure *pruning*; it reduces the computation time and makes the fit more robust. The pruning procedure is performed on each sample and region separately and the NP is only removed in those samples and regions where its effect is less than 1%. Studies were performed on the pruning threshold to verify that it has no impact on the final result.

In order to avoid statistical fluctuations in the fit we apply a *smoothing* procedure on the systematic uncertainties. This procedure averages systematic uncertainties across bins. The smoothing rebins the systematic variation distribution until the statistical uncertainty on each bin is below 8% of the number of events in that bin. If the derivative of the distribution changes sign four or less times (i.e. the distribution changes direction four or less times), the new binning is kept. If it changes direction more than four times, the statistical threshold is halved (i.e. 4% in the first iteration) and the process is repeated until the derivative sign variation equals four or less. The normalisation is kept fixed to the integral of the original distribution, thus the smoothing only affects the shape of the systematic.

7 FIT MODEL AND UNCERTAINTIES

In the $t\bar{t}H(H \to b\bar{b})$ analysis, a profile likelihood fit is performed in order to obtain the signal strength of the analysis given by $\mu_{t\bar{t}H} = \sigma_{\text{observed}}/\sigma_{\text{SM}}$. The fit model is constructed from all the variables used in each of the signal and control regions, the normalisation factors of specific processes, and the systematic uncertainties. These various components are discussed in this chapter. Before being applied to observed data, the model is first tested on the *Asimov* dataset (see section 6.3) in order to make sure it is well defined, flexible enough, and behaves properly.

7.1 Overview

The full fit is run over the nine signal regions and ten control regions of the $t\bar{t}H(H \to b\bar{b})$ analysis. The boosted channel has one dedicated signal region and the discriminant distribution fitted there is the boosted BDT output as shown in figure 5.28. The boosted channel is combined with all resolved signal and control regions in order to increase the sensitivity. In all signal regions of the resolved analysis, a dedicated classification BDT is used in the fit (see section 5.7.3). The binning of these BDT distributions is optimised in order to have maximum sensitivity and a low statistical uncertainty in each bin. The statistical uncertainties are minimised in order to avoid impacts on the final fit result from statistical fluctuations on the predicted event yield per bin.

The H_T^{had} distribution (the scalar sum of the jet p_T) is fitted in two of the resolved semileptonic control regions enriched in $t\bar{t}+\geq 1c$ (marked $\text{CR}^{5j}_{t\bar{t}+\geq 1c}$ and $\text{CR}^{\geq 6j}_{t\bar{t}+\geq 1c}$ in figure 5.12). This variable improves the signal sensitivity because it facilitates the distinction between the different components contributed by $t\bar{t}+$ light, $t\bar{t}+\geq 1c$, and $t\bar{t}+\geq 1b$. In all other resolved control regions, only one bin is fitted which represents the total event yield. This choice was made because the H_T^{had} variable is not well-modelled in these control regions and the mismodelling is not fully covered by the uncertainties.

It is observed that the data overshoots the predictions for $t\bar{t}$+HF. Therefore, the normalisations of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds relative to the total $t\bar{t}$ background are allowed to float freely in the fit. No prior knowledge is assumed about these normalisation factors (marked as $k(t\bar{t}+\geq 1b)$ and $k(t\bar{t}+\geq 1c)$) thus they are only constrained by the profile likelihood

fit to data. In all pre-fit results, the normalisation factors are set to one which corresponds to taking the prediction of the fractions of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ from the nominal POWHEG+ PYTHIA8 sample. The statistical and systematic uncertainties in the fit are discussed in more detail in the following sections.

7.2 Statistical uncertainties

The statistical uncertainties in the fit model are due to the finite number of both observed data events and simulated events. These uncertainties are summed in quadrature with the total systematic uncertainty since the two are uncorrelated. The statistical uncertainty on data per bin is calculated as the square root of the observed number of events in that bin (assuming a Gaussian distribution).

The statistical uncertainties on MC are included in the likelihood as additional NPs, one for each of the bins of the distributions fitted in the signal and control regions included in the fit. In general, we simulate more MC events than we observe in data which means that we need to apply an overall reweighting of the MC events in order to match the number of data events. This reweighting factor is < 1 and is the same for all events. The MC events are also weighted with other scale factors originating from their event generator, pile-up, b-tagging, and other sources. Since the MC events are weighted, a different approach to the error calculation needs to be taken. The standard deviation for N weighted events is defined as:

$$\sigma = \sqrt{\sum_{i=1}^{N} w_i^2}\,, \tag{7.1}$$

where w_i represents the total weight of the event i. The reweighting factor from MC to data is the same for all events and can therefore be taken out of the sum. Since this factor is less than one, the MC statistical uncertainties decrease with an increase in the number of generated events. The total uncertainty on each bin is computed as the quadratic sum of the MC statistical error of all background components. These uncertainties are added as nuisance parameters in the fit and treated as uncorrelated across all bins of the analysis.

7.3 Systematic uncertainties

The analysis is affected by many different sources of systematic uncertainties including experimental ones coming from the performance and simulation of the ATLAS detector and theoretical ones related to MC modelling and calculated cross-sections of the relevant processes. As discussed in section 6.5, NPs are assigned to each systematic uncertainty and included in the fit. Uncertainties are evaluated for each relevant sample in every region and can impact the normalisation of the sample and/or the shape of the final discriminants. Unless specified otherwise, the NPs are correlated across channels, analysis regions, and samples. In addition to the systematic uncertainties listed below, there are NPs in the fit for each bin in each of the

signal and control region distributions evaluating the statistical uncertainty of the simulated samples (see section 7.2).

7.3.1 Experimental uncertainties

All the sources of experimental uncertainties affect both the normalisation and the shape of the distributions in all samples, except for the luminosity uncertainty which only affects the normalisations. The uncertainty on the combined 2015 and 2016 luminosity is 2.1%. It is derived from dedicated van der Meer scans [121] performed at $\sqrt{s} = 13$ TeV in 2015 and 2016, following a method similar to that in reference [122]. The van der Meer method determines the effective transverse beam size by measuring the change in interaction rate when the two proton beams are slightly separated from their head-on collision configuration. The scans are performed in both the horizontal and vertical directions, using dedicated LHC fills with fewer bunches and lower intensities than the nominal conditions.

The adjustment of simulated data to match the pile-up distribution of observed data is another source of uncertainty. The uncertainty on the ratio of predicted and measured cross-sections of inelastic collisions is covered by this. The other sources of experimental systematic uncertainties are described in the following sections.

Jets

As described in section 4.3.1, the small jets used in the analysis are subject to a multi-stage calibration procedure that restores the energy scale of reconstructed jets to the scale of simulated truth jets. This JES calibration is a source of many systematic uncertainties in the analysis. The JES uncertainty per jet is small (in the order of a few percent) but its total effect is sizeable due to the large number of jets in the final state of this analysis. The full JES calibration yields 80 separate systematic uncertainties propagated from the individual calibrations, as detailed in reference [82]. However, a reduced set of NPs is available for analyses in order to simplify the implementation. This reduced set consists of 20 NPs coming from all stages of the JES calibration procedure and gives enough coverage of the uncertainties for most physics analyses.

Another factor of jet uncertainty comes from the jet energy resolution (JER). Due to noise, stochastic fluctuations in the calorimeter response, and detector calibration effects, we cannot measure the energy of a jet exactly. Rather, we measure jet energies along a Gaussian spread. The JER is defined as the width of the jet energy response distribution, $R_E = \langle E_{\text{reco}}/E_{\text{truth}} \rangle$ (whereas the JES is its mean), and its uncertainty is evaluated by smearing the energy of each jet in simulated events. The amount of smearing is determined from JER measurements in dijet, Z+jets, and γ+jets data events. The JER uncertainty is divided into two independent components: one for the semileptonic $\text{CR}^{5j}_{t\bar{t}+\geq 1c}$ and dileptonic $\text{CR}^{3j}_{t\bar{t}+\text{light}}$ regions, and one for all the other regions. This is done because the JER uncertainties in the first two regions show different behaviour from the other regions in the fit.

The last jet systematic comes from the uncertainty on the efficiency of jets passing the JVT cut (see section 4.3.3) that is meant to remove pile-up jets. All the uncertainties on the small jets are propagated directly to the large reclustered jets, as explained in section 4.4.4. This means that we do not have to take into account an additional set of large jet uncertainties.

Flavour tagging

The b-tagging used in the analysis (see section 4.3.4) has uncertainties on the tagging efficiency of b-jets as well as on the mistag rates of c-jets and light jets. These uncertainties are derived from the scale factors which correct for differences between data and MC simulated events. The b-jet tagging efficiency is measured in dileptonic $t\bar{t}$ data events. The mistag rate for c-jets is measured in $t\bar{t}$ events as well, identifying hadronically decaying W bosons including c-jets. The light jet mistag rate is measured in multijet events using jets whose secondary vertices and track impact parameters are consistent with a negative lifetime [89].

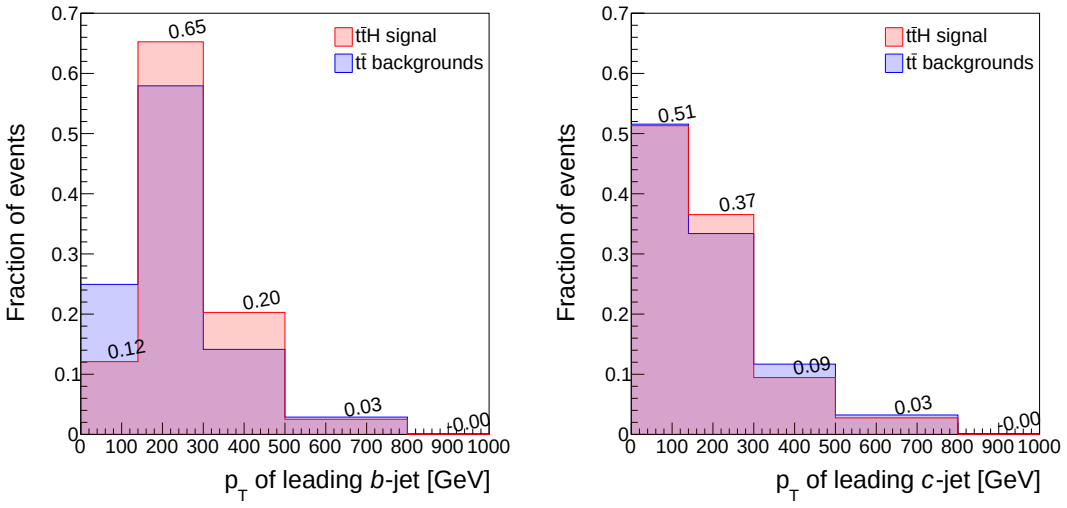
The efficiency and mistag rates are extracted for each of the four b-tagging working points and as a function of the jet kinematics. They are then combined into a distribution in which the correlations between the different WPs are taken into account. This leads to 30 independent uncertainties for the b-tagging efficiency, 15 for the c-jet mistag rate, and 80 for the light jet mistag rate. All uncertainties depend on the b-tagging working point as well as the jet p_T and range from 2 – 10% for b-tagging, 5 – 20% for c-mistagging, and 10 – 50% for light-mistagging. The mistag rate correction for c-jets and its uncertainty are also applied to τ-jets, with an additional uncertainty added to cover the extrapolation between c-jets and τ-jets.

Uncertainties associated with high-p_T heavy flavour jets

Since we are selecting high-p_T events in the boosted SR, it is expected that we have a significant fraction of high-p_T heavy flavour jets. It is therefore necessary to check what effect this has on our flavour tagging uncertainties. The ATLAS flavour tagging group advises that there may be additional uncertainties associated with b-jets of $p_T > 300$ GeV and c-jets of $p_T > 140$ GeV. Figure 7.1 shows the fraction of b-jets and c-jets with high transverse momentum in the boosted SR, for $t\bar{t}H$ signal events and $t\bar{t}$ background events. For signal events we have 23% of events with b-jets, and 49% of events with c-jets, above their respective high-p_T threshold. In background events, this is 17% for b-jets, and 48% for c-jets. This is a significant amount; therefore we need to study the effect of the extra uncertainties for high-p_T jets on the boosted BDT performance.

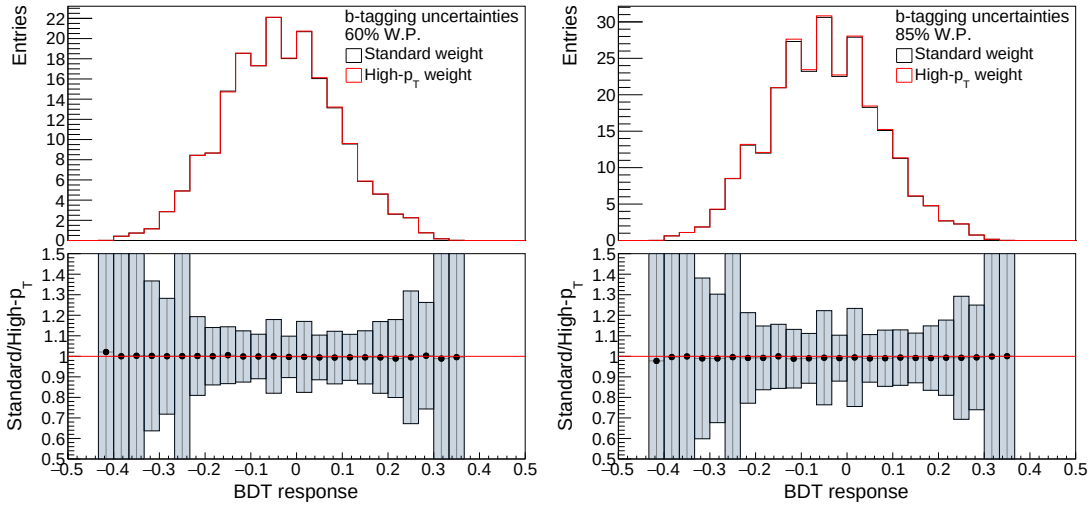
Currently, there is no high-p_T uncertainty calculation available for the continuous b-tagging scale that we use in our analysis. However, this uncertainty does exist for each of the four fixed b-tagging WPs. In figure 7.2 we show the effect of applying this high-p_T scale factor on our BDT response when using the 60% (a) and 85% (b) b-tagging WP. These plots show a negligible effect of this extra uncertainty factor on the boosted BDT. As a sanity check, figure 7.3 shows the effect of some other scale factors that we apply in the continuous b-tagging scen-

ario. These other uncertainties have a significantly larger effect on the BDT response than the high-p_T scale factors. Therefore, we conclude that the additional high-p_T uncertainties do not need to be applied in our SR.



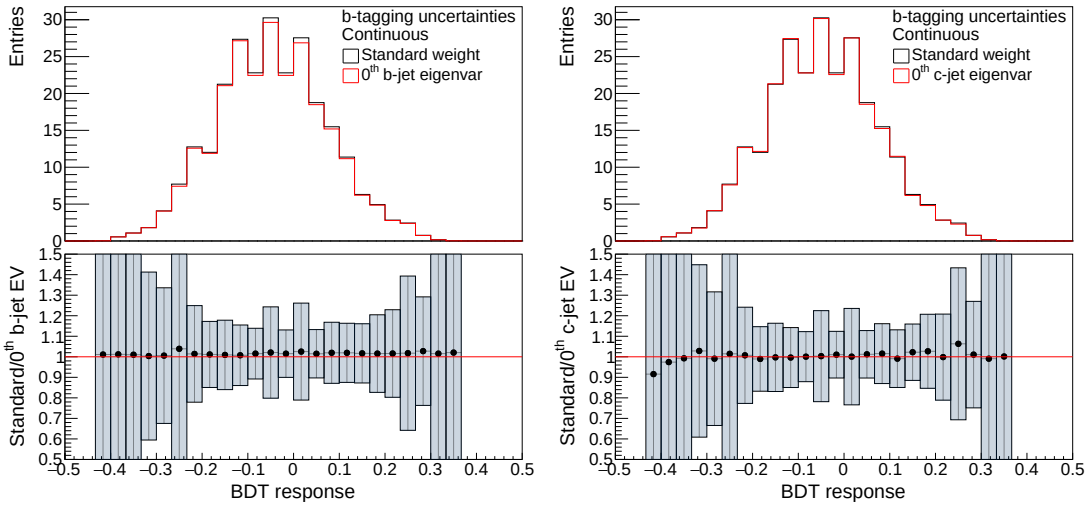
(a) (b)

Figure 7.1: The normalised p_T distributions in $t\bar{t}H$ signal and combined $t\bar{t}$ events for (a) the leading b-jet and (b) the leading c-jet in the boosted SR. The printed fractions correspond to the $t\bar{t}H$ signal. The flavour of jets is determined at MC truth level.



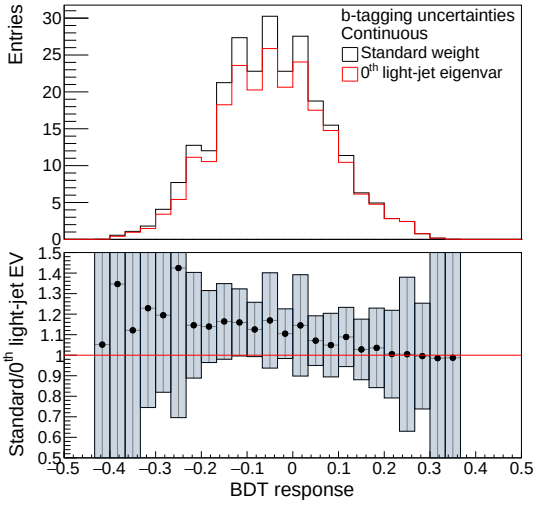
(a) (b)

Figure 7.2: The BDT response is shown using the full set of event weights with variations of the value used for the b-tagging scale factor. The standard scale factor is shown in black and the scale factor for high-p_T jets is shown in red. In (a) the scale factors for the b-tagging WP of 60% are shown, and (b) shows them for the 85% WP. The ratio plots show the difference in the BDT response between the standard and high-p_T scale factors.



(a)

(b)



(c)

Figure 7.3: The BDT response is shown using the full set of event weights with variations of the value used for the b-tagging scale factor. The standard scale factor is shown in black everywhere and the red line represents the scale factor for the 0^{th} eigenvar (one of the standard flavour tagging uncertainties described above). The ratio plots show the difference in the BDT response between the standard weight and the weight from the flavour tagging uncertainty described above. All comparisons are made for the continuous b-tagging scale factors.

Leptons

Lepton uncertainties arise from many sources, with the main ones being the efficiencies of the trigger, reconstruction, identification, and isolation. The lepton momentum scale and lepton momentum resolution also contribute to the lepton uncertainties. All uncertainties are defined as scale factors and measured by comparing data to simulation using leptons in various processes [68, 70]. There are 24 lepton uncertainties in total but they only have a very small impact on the results.

Missing transverse energy

The systematic uncertainties on the energy scales and energy resolutions of all physics objects are propagated to the MET.

7.3.2 Theoretical uncertainties

The theoretical predictions and models used in the analysis are sources of theoretical systematic uncertainties. The main uncertainties on a particular process come from the choice of the hard scatter model, parton shower model, and PDF set used for the event generation. These uncertainties are the source of the so-called *2-point systematics* because they are evaluated by comparing the nominal samples to alternative samples with a different configuration. The alternative samples are mostly generated using the AF2 fast detector simulation instead of the full GEANT4 simulation (see section 3.1.2). In these cases, an AF2 version of the nominal MC sample is used for the comparison such that there is no bias from the difference between the fullsim and fastsim. The difference between the two samples is defined as one standard deviation in the systematic uncertainty. This difference is applied as a 2-sided systematic uncertainty with the up variation one standard deviation above the nominal prediction and the down variation one standard deviation below the nominal prediction. We also include uncertainties on the theoretically calculated cross-sections and branching ratios of the processes in the analysis. The theoretical uncertainties related to modelling affect both the normalisation and shape of the distributions. The cross-section and normalisation uncertainties only affect the normalisation of the sample.

Signal modelling

The $t\bar{t}H$ cross-section uncertainty is split into two independent contributions. The first component is the QCD scale uncertainty evaluated as $^{+5.8\%}_{-9.2\%}$ [14]. The second component comes from the uncertainties on the PDF and α_s and is a symmetric $\pm 3.6\%$ [14]. The uncertainty on the branching fraction of the $H \to b\bar{b}$ decay mode is 2.2% [14]. The uncertainty on the choice of parton shower and hadronisation models is derived from a comparison between the nominal sample produced by MG5_aMC@NLO+PYTHIA8 and a sample produced with MG5_aMC@NLO interfaced to HERWIG++.

$t\bar{t}$ modelling

The $t\bar{t}H$ analysis is very sensitive to the $t\bar{t}$+ jets model used and it is therefore the process which has most systematic uncertainties assigned to it. As described in section 5.4, the MC generator used for the $t\bar{t}$+ jets nominal sample is POWHEG+PYTHIA8 which was found to describe data better than other available generators [109]. This nominal sample is compared to several other samples in order to evaluate the uncertainty on the choice of ME generator, PS and hadronisation models, and the scale settings. The full set of uncertainties applied to the $t\bar{t}$ background is listed in table 7.1. All uncertainties on the $t\bar{t}$ modelling, except the uncertainty on the cross-section, are evaluated independently for the $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$, and $t\bar{t}+$ light categories because these processes are affected by different types of uncertainties.

Systematic source	**Description**	**$t\bar{t}$ categories**
$t\bar{t}$ cross-section	Up or down by 6%	All, correlated
$k(t\bar{t}+\geq 1c)$	Free-floating $t\bar{t}+\geq 1c$ normalisation	$t\bar{t}+\geq 1c$
$k(t\bar{t}+\geq 1b)$	Free-floating $t\bar{t}+\geq 1b$ normalisation	$t\bar{t}+\geq 1b$
SHERPA5F vs. nominal	Choice of NLO event generator	All, uncorrelated
PS & hadronisation	POWHEG+HERWIG7 vs. POWHEG+PYTHIA8	All, uncorrelated
ISR/FSR	Variations of μ_R, μ_F, h_{damp} and A14 Var3c parameters	All, uncorrelated
$t\bar{t}+\geq 1c$ ME vs. inclusive	MG5_aMC@NLO+HERWIG++: ME prediction (3F) vs. incl. (5F)	$t\bar{t}+\geq 1c$
$t\bar{t}+\geq 1b$ SHERPA4F vs. nominal	Comparison of $t\bar{t}+b\bar{b}$ NLO (4F) vs. POWHEG+PYTHIA8 (5F)	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ renorm. scale	Up or down by a factor of two	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ resumm. scale	Vary μ_Q from $H_T/2$ to μ_{CMMPS}	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ global scales	Set μ_Q, μ_R, and μ_F to μ_{CMMPS}	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ shower recoil scheme	Alternative model scheme	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ PDF (MSTW)	MSTW vs. CT10	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ PDF (NNPDF)	NNPDF vs. CT10	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ UE	Alternative set of tuned parameters for the UE	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 1b$ MPI	Up or down by 50%	$t\bar{t}+\geq 1b$
$t\bar{t}+\geq 3b$ normalisation	Up or down by 50%	$t\bar{t}+\geq 1b$

Table 7.1: Summary of the systematic uncertainties on the $t\bar{t}$+jets backgrounds. The last column indicates the $t\bar{t}$ category to which the systematic uncertainties are applied and in the case where it is applied to all categories, whether it is considered as correlated or uncorrelated across them.

The first section in table 7.1 shows the uncertainty on the inclusive $t\bar{t}$ cross-section, evaluated at ±6% [110]. The normalisations for both $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ are left free-floating in the fit because these processes have relatively weak constraints from data measurements. Studies carried out during the Run I analysis have shown that MC simulations underestimate these processes in comparison to data, as much as 40% for the $t\bar{t}+b\bar{b}$ subcomponent [25]. The normalisations are set to one in all pre-fit distributions and calculations.

The second section in table 7.1 lists the various systematics covering the uncertainties in the modelling choices of all $t\bar{t}$ categories. The fractions of $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$, and $t\bar{t}+$ light in the

alternative samples used to evaluate these systematics are reweighted to match the POWHEG+PYTHIA8 prediction. The various subcategories of the $t\bar{t}+\geq 1b$ component ($t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$, and $t\bar{t}+\geq 3b$) are reweighted to match the SHERPA4F prediction as discussed in section 5.4.2. These efforts make sure that the shape of the uncertainties derived are not affected by the different relative fractions of the various $t\bar{t}$ (sub)categories between samples, avoiding double counting of the normalisation uncertainties.

In order to evaluate the uncertainty coming from the choice of the ME event generator, the nominal POWHEG+PYTHIA8 sample is compared to a SHERPA2.2.1 sample. Note that this changes both the ME model and the PS and hadronisation models. The alternative would be to compare POWHEG+ PYTHIA8 to MG5_aMC@NLO+PYTHIA8, but as we saw in section 5.7.2, the MG5_aMC@NLO generator leads to a large number of negatively weighted events which means a low effective number of events. We therefore chose the SHERPA sample because it has better statistics for the comparison. The uncertainty on the PS and hadronisation models is estimated by comparing the nominal POWHEG+PYTHIA8 sample to a POWHEG+ HERWIG7 sample.

The ISR/FSR uncertainty assesses the impact of additional radiation in $t\bar{t}$ events. This is done with two alternative POWHEG+ PYTHIA8 $t\bar{t}$ samples in which the h_{damp} parameter, renormalisation scale (μ_R), and factorisation scale (μ_F) are varied [109]. The h_{damp} parameter controls the ME-to-PS matching in POWHEG and regulates the damping of high-p_T radiation. The alternative samples also change the A14 tune using the Var3c variation which covers the size of other available tuning configurations for this sample. One of the alternate samples increases the amount of radiation by using the Var3c *up* variation, decreasing μ_R and μ_F by a factor two, and increasing h_{damp} by a factor two. The other alternate sample decreases the amount of radiation by using the Var3c *down* variation, increasing μ_R and μ_F by a factor two, and keeping h_{damp} fixed at its nominal value of $1.5m_{\text{top}}$.

In the $t\bar{t}+\geq 1c$ background modelling, one needs to choose an approach in which the charm jets are either mostly produced in the parton shower, or mostly in the hard scatter. Since it is unclear from theory or experiment which of these approaches is more accurate, we add a systematic evaluating this uncertainty. It is derived from the comparison between a NLO prediction with $t\bar{t}+c\bar{c}$ in the ME including massive c-quarks (a three-flavour scheme PDF) and an inclusive sample using a five-flavour scheme in the PDF with massless c-quarks. In the latter sample, the $t\bar{t}+\geq 1c$ process originates only in the PS. Both samples are produced with MG5_aMC@NLO+HERWIG++. The difference between the two samples is applied as an independent systematic on the $t\bar{t}+\geq 1c$ background in the nominal sample.

The last uncertainty in the second section of table 7.1 covers the difference between the nominal $t\bar{t}+\geq 1b$ description and the SHERPA4F description. The nominal sample is an inclusive sample with a five-flavour scheme with massless b-quarks, whereas the SHERPA4F sample has $t\bar{t}+b\bar{b}$ in the ME including massive b-quarks (four-flavour scheme). Since the relative fractions of the various $t\bar{t}+\geq 1b$ subcategories are reweighted from the nominal to the SHERPA4F prediction, this systematic only accounts for the difference in shape between the two samples. The uncertainty is not applied to the $t\bar{t}+b$ (MPI/FSR) subcategory since this is not included in

the four-flavour calculation.

The third section in table 7.1 lists systematic uncertainties that only affect the fractions of the various $t\bar{t}+\geq 1b$ subcategories. These fractions are all fixed to the SHERPA4F prediction and seven systematic uncertainties are applied in order to cover the uncertainty on this prediction. Three of these systematics are evaluated by varying the scales in the SHERPA4F sample: the renormalisation scale is decreased and increased by a factor of two, the resummation scale is set from $H_T/2$ to μ_{CMMPS}, and a global scale choice is set as $\mu_Q = \mu_F = \mu_R = \mu_{\text{CMMPS}}$. Two other systematics cover the difference between the nominal CT10 PDF set [41, 42] and two alternatives: MSTW2008NLO [123] and NNPDF2.3NLO [39]. One systematic is evaluated by choosing an alternative shower recoil scheme and the last one by using an alternative set of tuned parameters for the UE. These seven uncertainties form the uncertainty band on the SHERPA4F prediction shown in figure 5.4.

Two additional systematic uncertainties are applied to the $t\bar{t}+\geq 1b$ background normalisations. Firstly, an extra 50% normalisation uncertainty is added for the $t\bar{t}+\geq 3b$ subcategory because there is a large difference between the nominal and SHERPA4F predictions which is not covered by the uncertainties described above. A 50% normalisation uncertainty is also applied to the $t\bar{t}+\geq 1b$ events from MPI which is based on studies of different tunable sets of parameters for the UE calculation. Note that variations related to the fraction and shape of the $t\bar{t}+b$ (MPI/FSR) subcategory are already incorporated in the uncertainties described above, since the fraction of this subcategory is not fixed to the SHERPA4F prediction like the other $t\bar{t}+\geq 1b$ subcategories.

Other background modelling

For the other backgrounds in the analysis, the main sources of uncertainty come from the theoretical uncertainty on the production cross-sections. All systematics considered for the background samples, excluding $t\bar{t}+$ jets, are summarised in table 7.2.

The uncertainty on the choice of ME generator, PS model, and hadronisation model of the $t\bar{t}+V$ background is extracted from a comparison between the nominal MG5_aMC@NLO+PYTHIA8 sample and a SHERPA sample. The uncertainties on $t\bar{t}W$ and $t\bar{t}Z$ are treated as uncorrelated.

The uncertainties on the W+jets and Z+jets samples are derived from variations of μ_F, μ_R, and matching parameters in the nominal SHERPA sample. For Z+jets, the normalisation uncertainty is uncorrelated across jet bins.

For the Wt and t-channel single-top samples, an uncertainty on the choice of PS and hadronisation model is evaluated from a comparison between the nominal POWHEG+PYTHIA6 sample and two alternative samples. One of the alternative samples is generated using POWHEG+ HERWIG++ and the other uses POWHEG+PYTHIA6 but varies the μ_F, μ_R, and appropriate Perugia 2012 tuning parameters. An extra uncertainty is added to the Wt sample which covers the difference between the diagram removal (used as default) and diagram subtraction schemes. These are schemes that remove the interference effects between the overlapping $t\bar{t}$ and Wt

events.

The uncertainty on the estimated yield of the fakes is 50% in the semileptonic channel and 25% in the dileptonic channel. In the semileptonic case, the uncertainty is decorrelated for the e and μ channels, between the resolved and boosted categories, and between the resolved regions with 5 jets and ≥ 6 jets. In the dileptonic channel the uncertainty is correlated across all regions and lepton flavours.

The other entries in table 7.2 refer to the uncertainties in the cross-section calculation. For some samples, these are split into a QCD scale (μ_F and μ_R) uncertainty and an uncertainty on the PDF used in the event generation.

Process	**Type**	**Systematic description**
$t\bar{t}+V$	N	15% cross-section uncertainty
	S&N	ME, PS, and hadronisation
W+jets	N	40% cross-section uncertainty
	N	30% W+2 HF-jets normalisation uncertainty
	N	30% $W+ \geq 3$ HF-jets normalisation uncertainty
Z+jets	N	35% normalisation uncertainty
Single top: Wt-channel	N	$^{+5\%}_{-4\%}$ cross-section uncertainty
	S&N	PS and hadronisation
	S&N	Diagram removal vs. diagram subtraction schemes
Single top: t-channel	N	$^{+5\%}_{-4\%}$ cross-section uncertainty
	S&N	PS and hadronisation
Single-top: s-channel	N	$^{+5\%}_{-4\%}$ cross-section uncertainty
$tHjb$	N	QCD$^{+6.5\%}_{-14.9\%}$ and PDF ±3.7%
tWH	N	QCD$^{+6.5\%}_{-6.7\%}$ and PDF ±6.3%
$t\bar{t}WW$	N	QCD$^{+10.9\%}_{-11.8\%}$ and PDF ±2.1%
tZ	N	50% cross-section uncertainty
tZW	N	50% cross-section uncertainty
4-top	N	50% cross-section uncertainty
Diboson	N	50% cross-section uncertainty
Fakes	N	50% (25%) normalisation uncertainty

Table 7.2: Summary of the systematic uncertainties on the background samples, excluding $t\bar{t}$+jets. The second column indicates whether the uncertainty affects the sample normalisation (N) or shape (S). The uncertainty on the fakes is given for the semileptonic and dileptonic (in brackets) channels separately.

8 RESULTS

The results of the $t\bar{t}H(H \to b\bar{b})$ analysis are presented in this chapter. The results of the boosted analysis are given in section 8.1. The combination of the boosted and resolved channels is discussed in section 8.2. The results of a combination of this analysis with other searches for $t\bar{t}H$ in the ATLAS collaboration, which are optimised for other Higgs boson decay modes, are detailed in section 8.3. The current search builds on previous searches for the same process performed with ATLAS data recorded at $\sqrt{s} = 7$ TeV [24] and 8 TeV [25, 26]. The signal strength found for the full combination of Run I $t\bar{t}H(H \to b\bar{b})$ searches in ATLAS is $\mu_{t\bar{t}H} = 1.4 \pm 1.0$ [26].

8.1 Boosted analysis results

The boosted analysis is designed to be combined with the resolved analysis because it is currently not sensitive enough on its own. A fit over just the boosted signal region is performed in order to check the current performance and get an estimate of its impact in the future. The signal strength of $t\bar{t}H$ is extracted using the methods described in chapter 6, with all uncertainties discussed in chapter 7 included. The extracted best-fit value for the signal strength parameter μ is

$$\mu_{t\bar{t}H} = 1.46^{+2.70}_{-2.80}, \tag{8.1}$$

for a Higgs mass of 125 GeV. The upper limit on μ is computed using the CL_s method (see section 6.4). A signal strength larger than 6.4 is excluded at the 95% confidence level. The main uncertainties impacting the boosted result are the normalisation factors on the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds, as well as the uncertainty on the $t\bar{t}+\geq 1b$ PS and hadronisation model. The latter is evaluated by comparing the nominal POWHEG+PYTHIA8 sample to a POWHEG+HERWIG7 sample. As expected, the sensitivity of the boosted channel alone is currently very low. The large dataset at the end of Run II and beyond will allow for a more significant contribution of this region. In the future, the boosted region can be used for a differential cross-section measurement in the high-p_T phase space region.

8.2 Combination of boosted and resolved $t\bar{t}H(H \to b\bar{b})$ channels

The full combination of the boosted and resolved analyses is discussed here. The signal strength of $t\bar{t}H$ is extracted using the methods described in chapter 6. A binned profile likelihood fit is performed over all resolved and boosted regions simultaneously and the systematic uncertainties discussed in section 7.3 enter the fit as NPs. No distinction is made between signal and control regions in the fit, except for the discriminant variable used. The results presented here are obtained using the RooFit framework [124] with the Minuit2 package [125] for the determination of the best-fit values of the signal strength and the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisation factors.

8.2.1 Fit to Asimov data

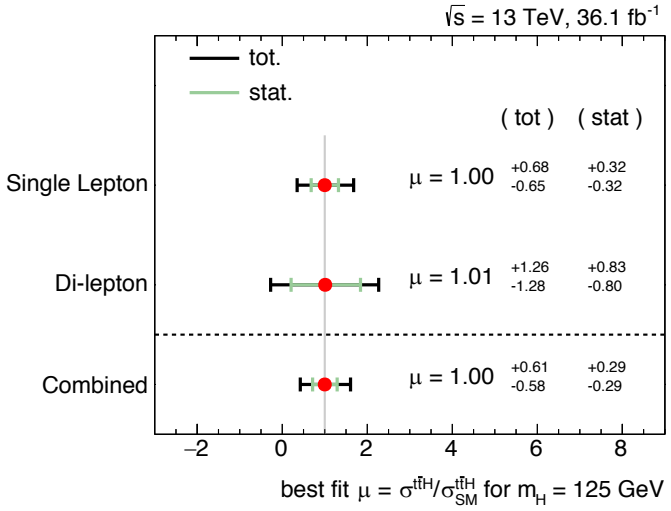
Before the fit is applied to observed data, the model is validated on the Asimov dataset. This dataset is constructed from the nominal predicted values of all parameters in the fit model. A Poisson error corresponding to the statistical uncertainty of the data is assumed in each bin. The expected combined signal strength found is

$$\mu_{\text{Asimov}} = 1.00 \pm 0.29(\text{stat.})^{+0.54}_{-0.50}(\text{syst.}) = 1.00^{+0.61}_{-0.58}, \tag{8.2}$$

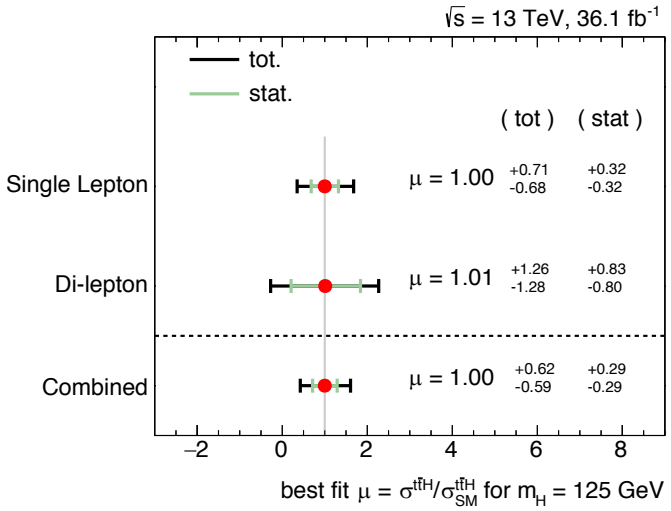
corresponding to an expected significance of 1.6σ. The Asimov signal strength in the single-lepton channel alone is $1.00 \pm 0.32(\text{stat.})^{+0.60}_{-0.57}(\text{syst.}) = 1.00^{+0.68}_{-0.65}$, corresponding to an expected significance of 1.5σ. The statistical uncertainty is obtained by fixing all the NPs to their post-fit values (except for the two normalisation factors and μ itself) and redoing the fit. The systematic component of the total uncertainty is then calculated by subtracting the statistical component from the total in quadrature. Figure 8.1(a) shows the signal strengths for the combined as well as the individual single-lepton and dilepton fits to the Asimov dataset.

The Asimov fit is run on the full combination of all single-lepton and dilepton regions, including the boosted signal region. In order to check the final sensitivity of the boosted region on the combined analysis, an Asimov fit is also run when this region is excluded. In this case, any events that fell into the boosted signal region and were overlapping with any of the resolved regions, are now given to the resolved analysis. The results of this fit are shown in figure 8.1(b). The semileptonic signal strength is now $\mu_{\text{Asimov}} = 1.00 \pm 0.32(\text{stat.})^{+0.63}_{-0.60}(\text{syst.}) = 1.00^{+0.71}_{-0.68}$. The addition of the boosted region thus slightly reduces the systematic uncertainty of the semilepton analysis. However, the impact on the full combined fit is very small, with the total uncertainty brought down from $^{+0.62}_{-0.59}$ to $^{+0.61}_{-0.58}$.

The distributions of the eight input variables to the boosted BDT are shown in figures 8.2 through 8.5 both before and after the single-lepton fit to the Asimov dataset. As expected, we see a clear reduction in the post-fit uncertainties compared to the pre-fit level. This is due to the generation of constraints on the NPs, as well as correlations between them, by the fit to Asimov data.



(a) Full combined fit including the boosted signal region



(b) Combined fit excluding the boosted signal region

Figure 8.1: Signal strength μ from a fit to Asimov data for the individual semileptonic and dileptonic channels, as well as for the combination. Figure (a) includes the boosted signal region whereas it is excluded in the fit results in figure (b).

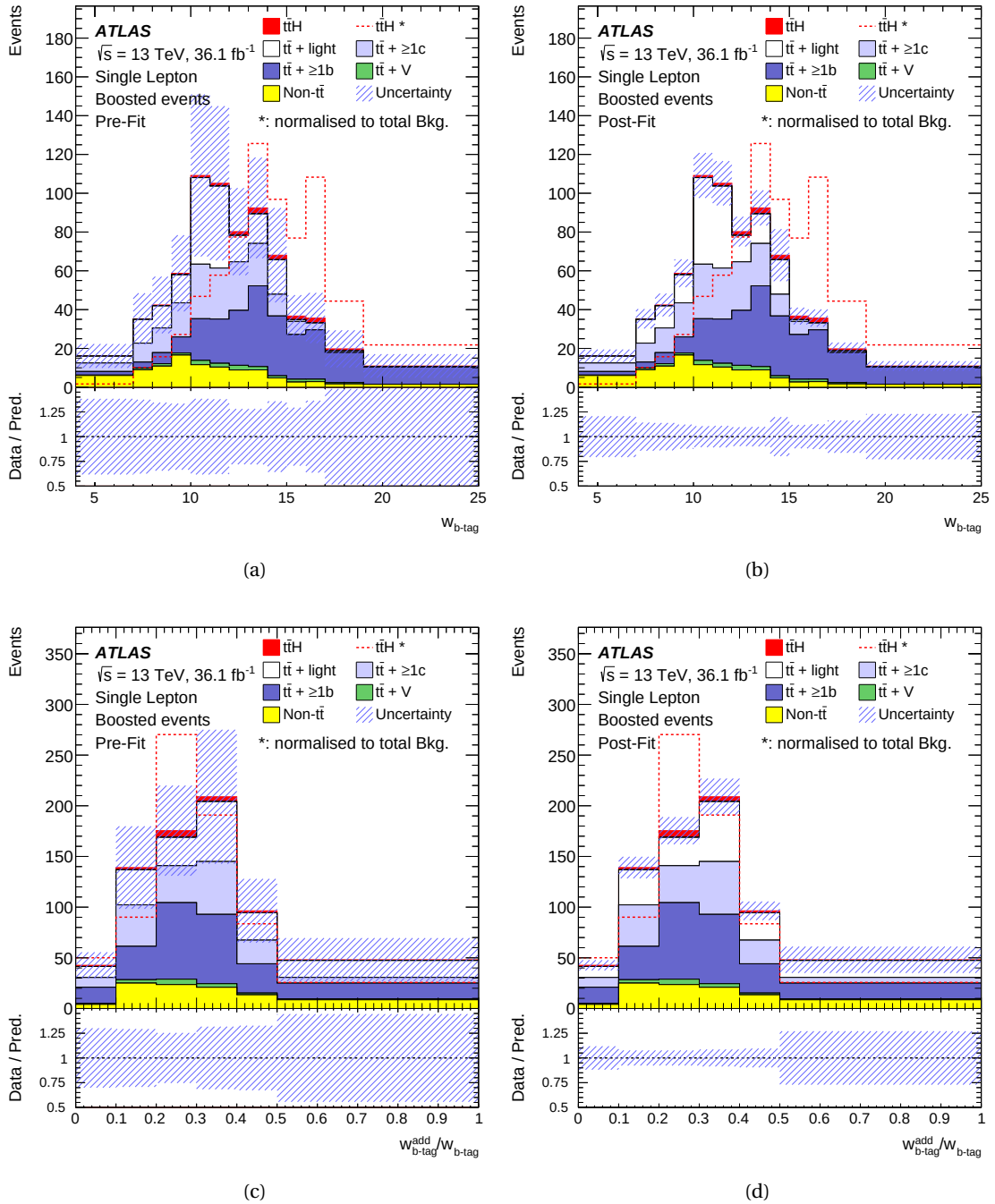


Figure 8.2: The two b-tagging variables used in the boosted classification BDT before (left) and after (right) the single-lepton Asimov fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.

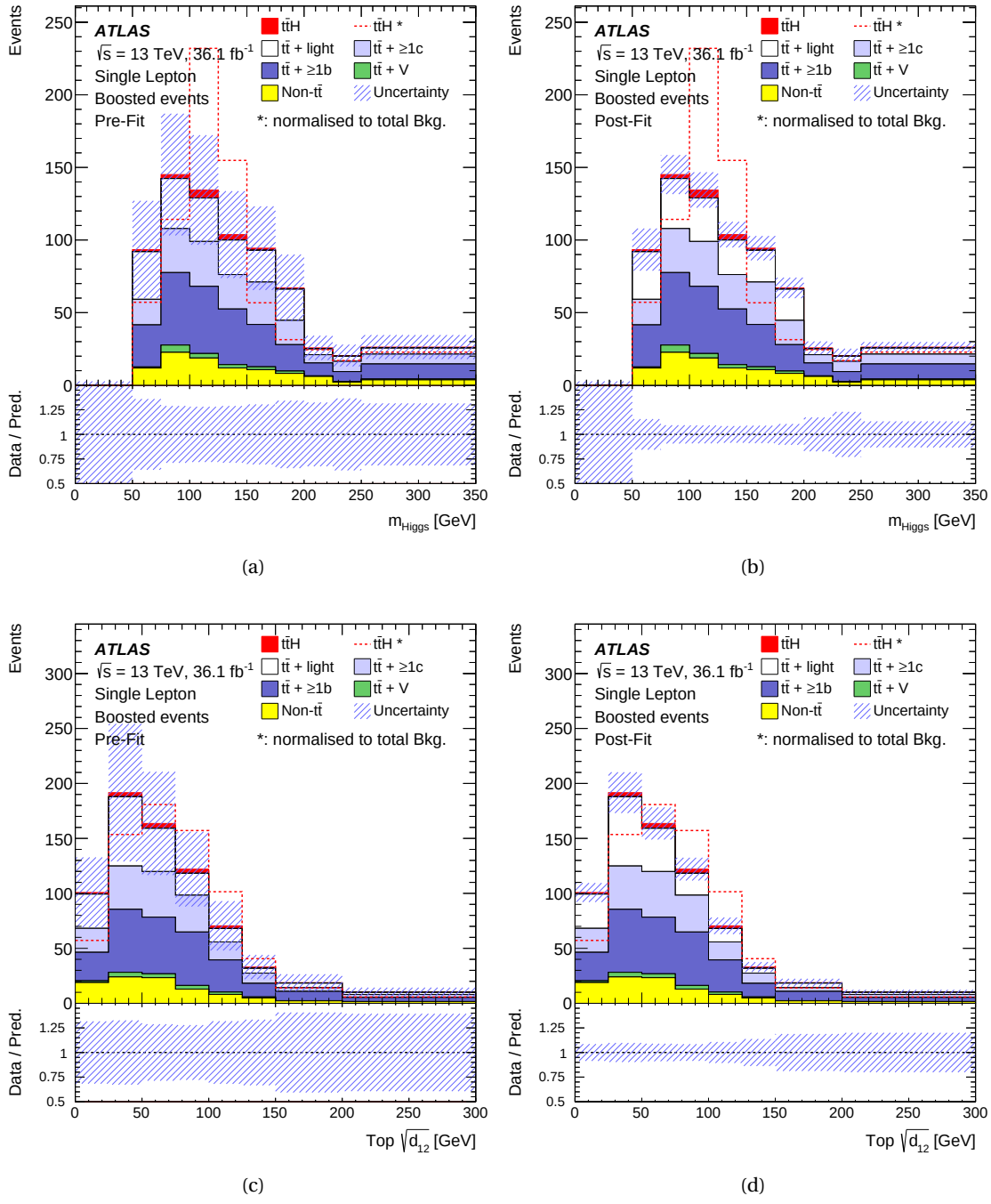


Figure 8.3: The two jet substructure variables used in the boosted classification BDT before (left) and after (right) the single-lepton Asimov fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.

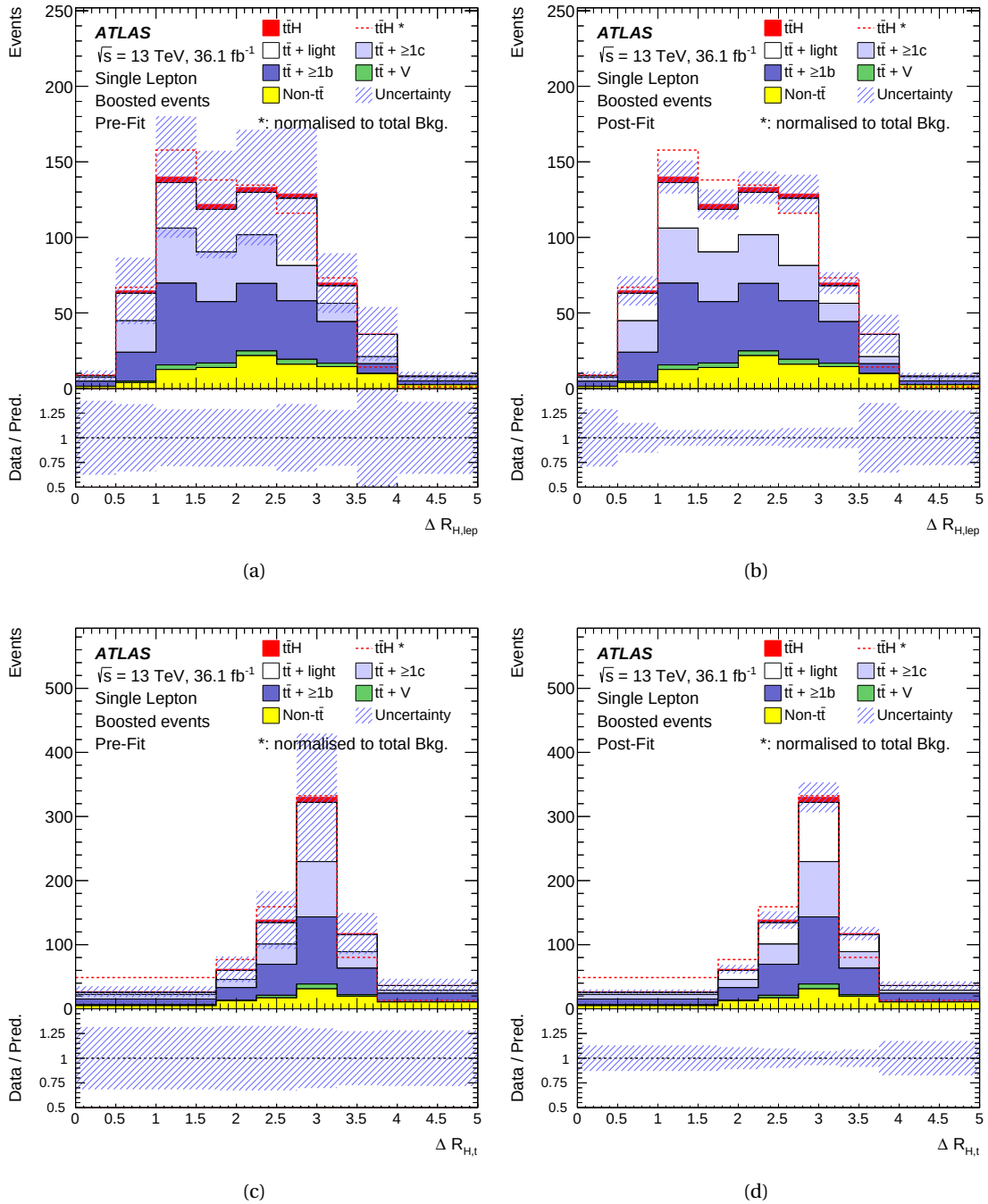
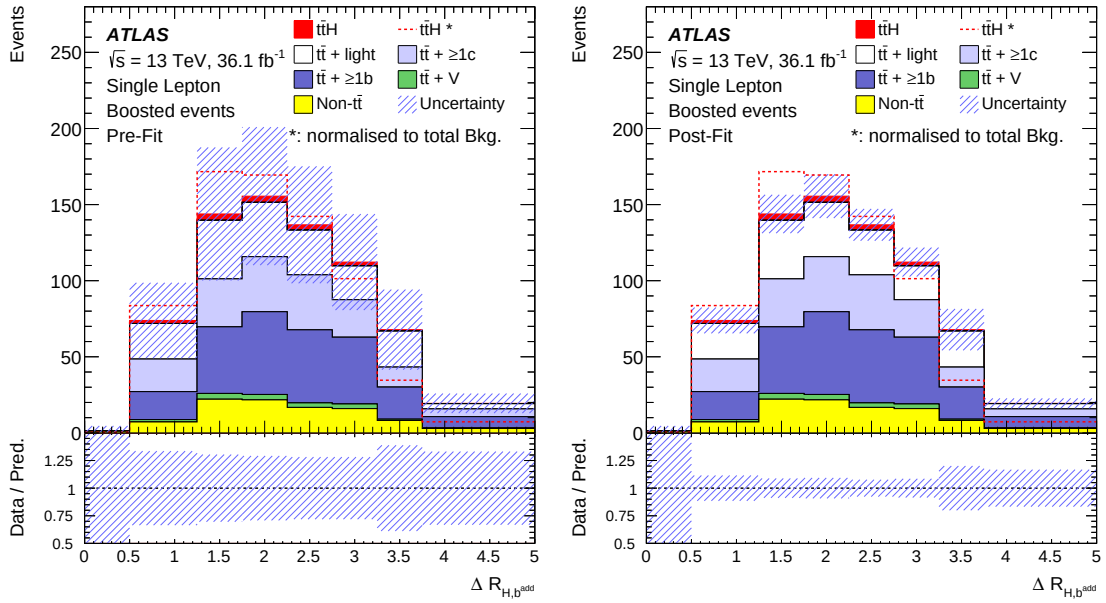
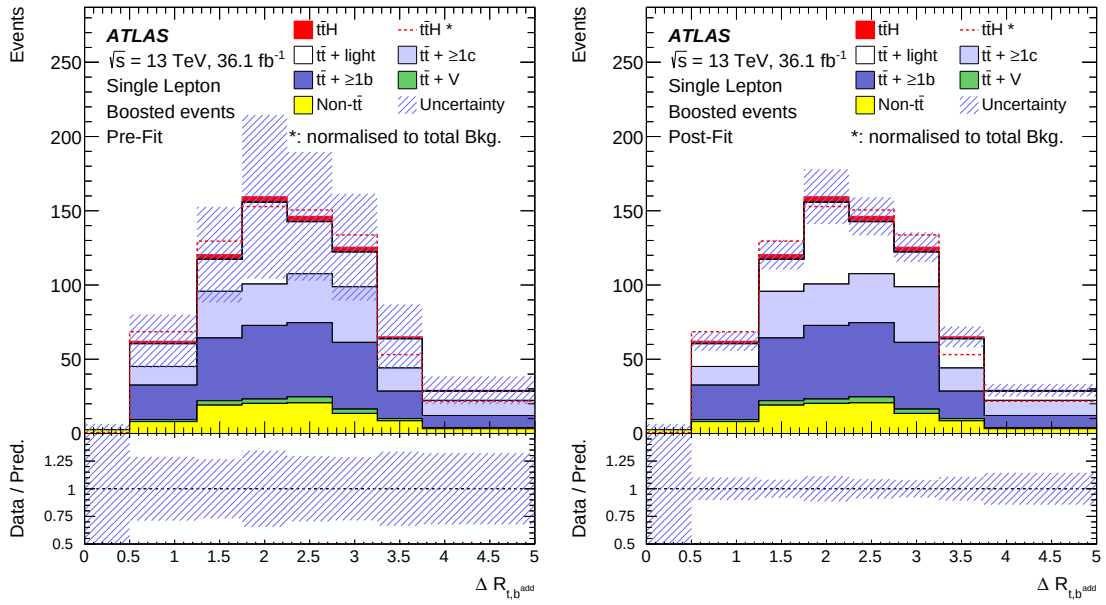


Figure 8.4: Two of the angular variables used in the boosted classification BDT before (left) and after (right) the single-lepton Asimov fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.



(a)

(b)



(c)

(d)

Figure 8.5: Two of the angular variables used in the boosted classification BDT before (left) and after (right) the single-lepton Asimov fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.

Figure 8.6 shows the twenty systematic uncertainties with the highest impact on the signal strength in the single-lepton Asimov fit, including the boosted signal region. By construction, all the NPs are centred around zero and the $t\bar{t}+\geq 1b$ normalisation factor, k, is centred at one. We observe some constraints on various NPs in the fit, i.e. the post-fit uncertainty is smaller than the pre-fit uncertainty and therefore $|(\hat{\theta}-\theta_0)/\Delta\theta| < 1$. This shows that the fit is capable of constraining the systematic variations which would otherwise lead to significant deviations from data in the discriminant distributions.

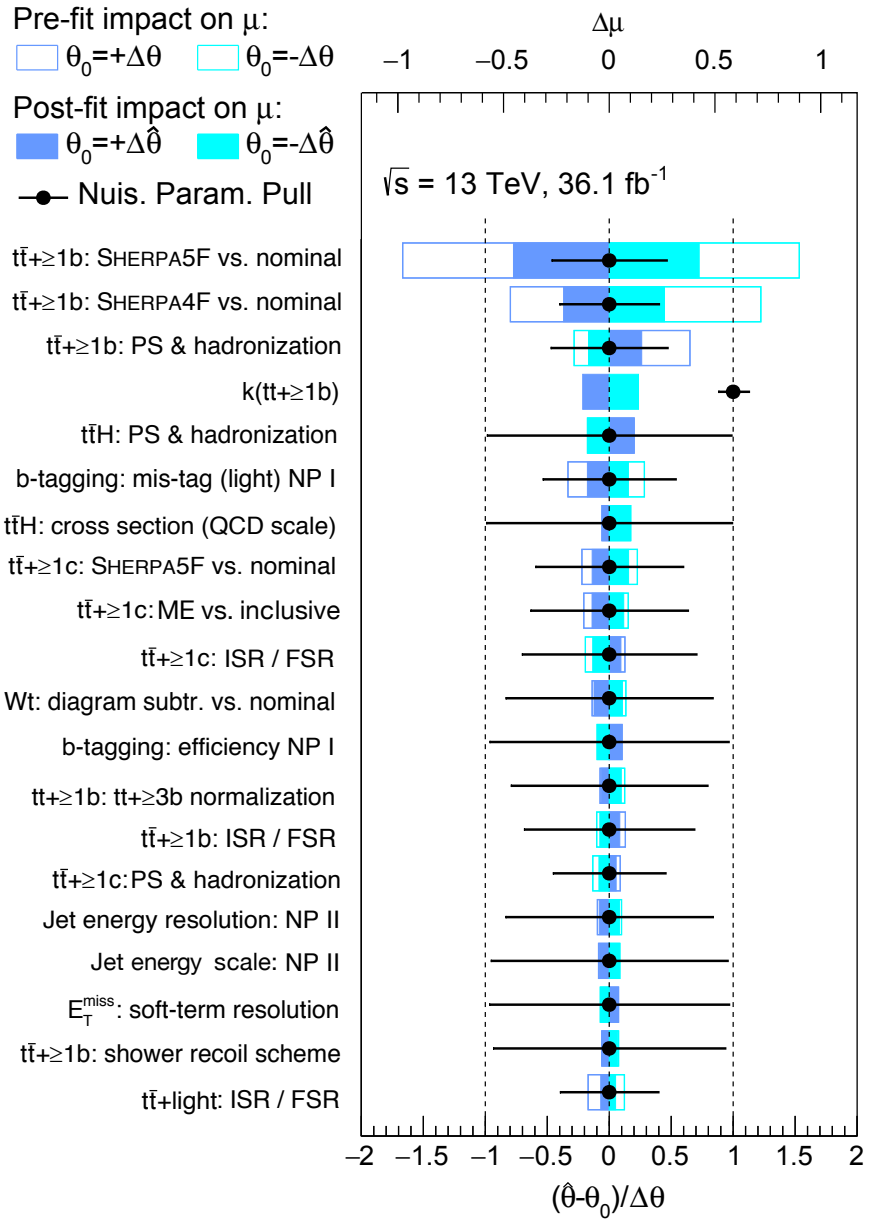


Figure 8.6: The top 20 nuisance parameters in the semileptonic fit (including the boosted signal region) to Asimov data, ranked according to their impact ($\Delta\mu$) on the measured signal strength μ. Both the pre-fit (empty blue rectangles) and post-fit (filled blue rectangles) impact on μ are shown and correspond to the top axis. The black dots denote the pulls of the nuisance parameters from their nominal value θ_0 with the black bar indicating the post-fit error relative to the pre-fit error, after applying the constraints from the fit. The pulls and constraints correspond to the bottom axis.

The first four NPs are all related to the $t\bar{t}+\geq 1b$ background (see table 7.1). In fact, the modelling of $t\bar{t}+\geq 1b$ contributes $^{+0.49}_{-0.48}$ to the total systematic uncertainty of the single-lepton Asimov fit. In addition, the $t\bar{t}+\geq 1b$ normalisation contributes an uncertainty of $^{+0.12}_{-0.14}$. These numbers are obtained by fixing all the $t\bar{t}+\geq 1b$ uncertainties, repeating the fit, and then subtracting in quadrature the resulting uncertainty from the total uncertainty of the full fit. The modelling of the dominant $t\bar{t}+\geq 1b$ background is thus clearly the main source of uncertainty for the single-lepton channel in this analysis.

The systematic ranking in figure 8.6 also tells us that the uncertainties on the signal modelling are non-negligible. In particular, the fifth ranked systematic covers the uncertainty in the parton shower and hadronisation model of the $t\bar{t}H$ signal sample, which is evaluated by comparing PYTHIA8 to HERWIG++. However, the impact of the signal modelling systematics are still sub-dominant compared to those on $t\bar{t}+\geq 1b$ and are not constrained.

8.2.2 Fit to pseudo data

A fit to pseudo data is performed to test the robustness of the fit with respect to the choice of $t\bar{t}$+jets model. The pseudo dataset is built from the Asimov dataset in which the nominal $t\bar{t}$ sample modelled by POWHEG+ PYTHIA8 is replaced by a POWHEG+ PYTHIA6 sample. This sample is not used in the definition of any uncertainty and was generated in the same way as for the Run I analysis in [25]. The relative fractions of the $t\bar{t}+\geq 1b$ subcategories in the POWHEG+ PYTHIA6 sample are reweighted to the SHERPA4F prediction, just as for the nominal $t\bar{t}$ sample.

The fit to pseudo data is performed to test the fit model's accommodation for mismodelling in the $t\bar{t}$ background. If the systematic uncertainties on $t\bar{t}$ are sufficiently flexible to cover the difference between the nominal model and the pseudo data model, only these $t\bar{t}$ uncertainties should be pulled whereas all other systematics should remain at their nominal value. The signal strength should also remain unchanged. This alternative fit showed no bias in the signal extraction and is able to recover the difference in modelling by using the $t\bar{t}$-related systematic uncertainties. This gives extra confidence in the robustness of the fit model.

8.2.3 Agreement between data and prediction

After the checks described above, the fit to measured data is performed. In order to make sure that our simulation model describes the data accurately, we test the agreement between the data and predicted simulation in several distributions.

The predicted and observed event yields for all signal and control regions are shown in figure 8.7. The predicted event yields are shown before the fit to data (pre-fit) and after the fit to data (post-fit) under the signal-plus-background hypothesis. The normalisation factors for the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ processes are set to one for the pre-fit results. The pre-fit plots therefore do not include any uncertainty for the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations. The yields for each region show a reasonable agreement between data and simulated events pre-fit within the

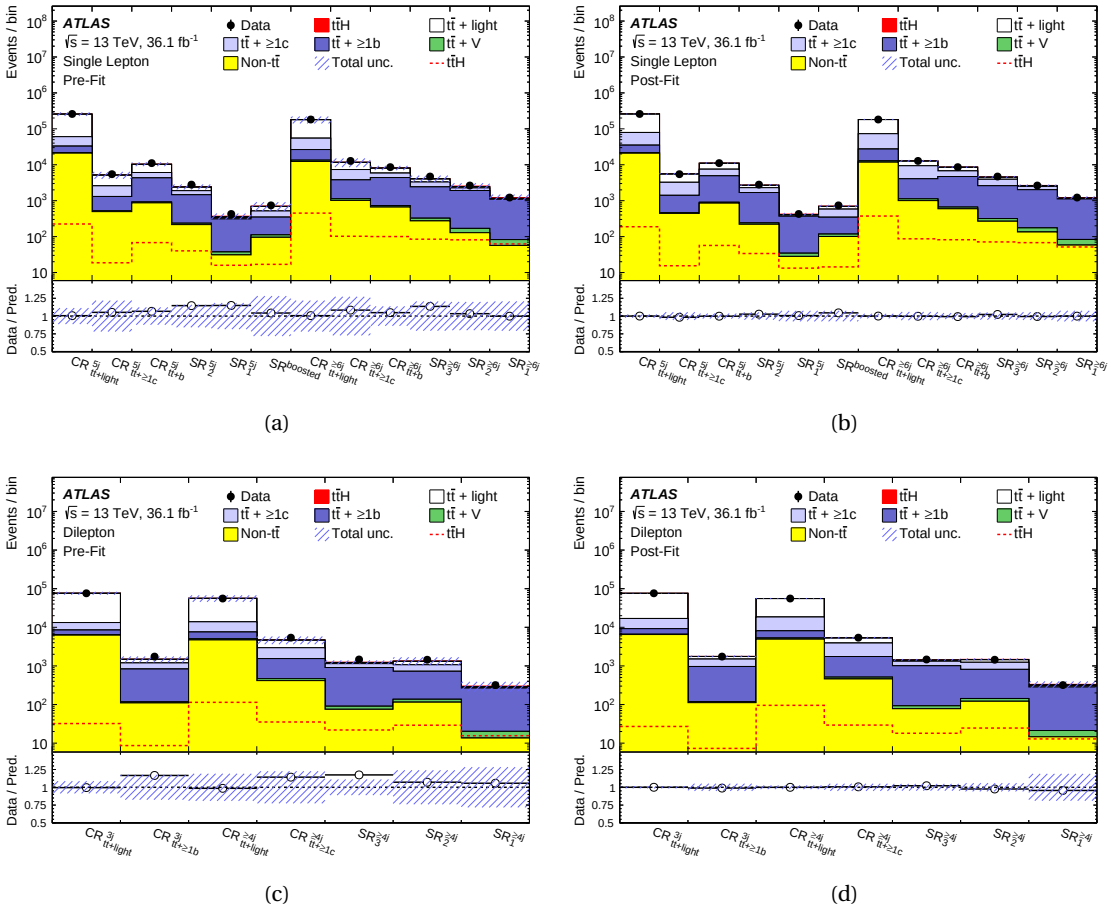


Figure 8.7: Event yields of observed data and predicted simulation for all signal and control regions [103]. The semileptonic channels are shown pre-fit in (a) and post-fit in (b), dileptonic channels are pre-fit in (c) and post-fit in (d). The $t\bar{t}H$ signal is shown in red, both stacked on top of the backgrounds and in a dashed line for better visibility, normalised to the SM cross-section pre-fit and to the fitted μ post-fit.

total uncertainty indicated by the blue hashed area. The agreement is improved significantly in the post-fit plots due to the NPs being adjusted from their nominal value by the fit. The normalisations of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ are an example thereof, with their best-fit values resulting in 1.24 ± 0.10 and 1.63 ± 0.23, respectively. Note that the uncertainties quoted on these measured normalisation factors do not include the theory uncertainty on the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ cross-sections. The post-fit uncertainty in figures 8.7 (b) and (d) is reduced because the fit constrains the NPs and generates correlations between them. The good agreement between data and simulation within the post-fit uncertainties gives confidence in the validity of the extrapolation of constraints and pulls across the various analysis regions.

Figures 8.8 through 8.11 show the distributions of the eight input variables to the boosted BDT before and after performing the single-lepton fit to data. The post-fit distributions show a slight improvement in the data-simulation agreement compared to the pre-fit ones. The post-fit uncertainties are significantly reduced. This reduction in post-fit uncertainties is very similar to that shown in figures 8.2 – 8.5 for the fit to Asimov data. Some small deviations of prediction from data are observed but none large enough to cause concern. There is no discernible trend in the modelling of the shape of the distributions or clear offset in the normalisations.

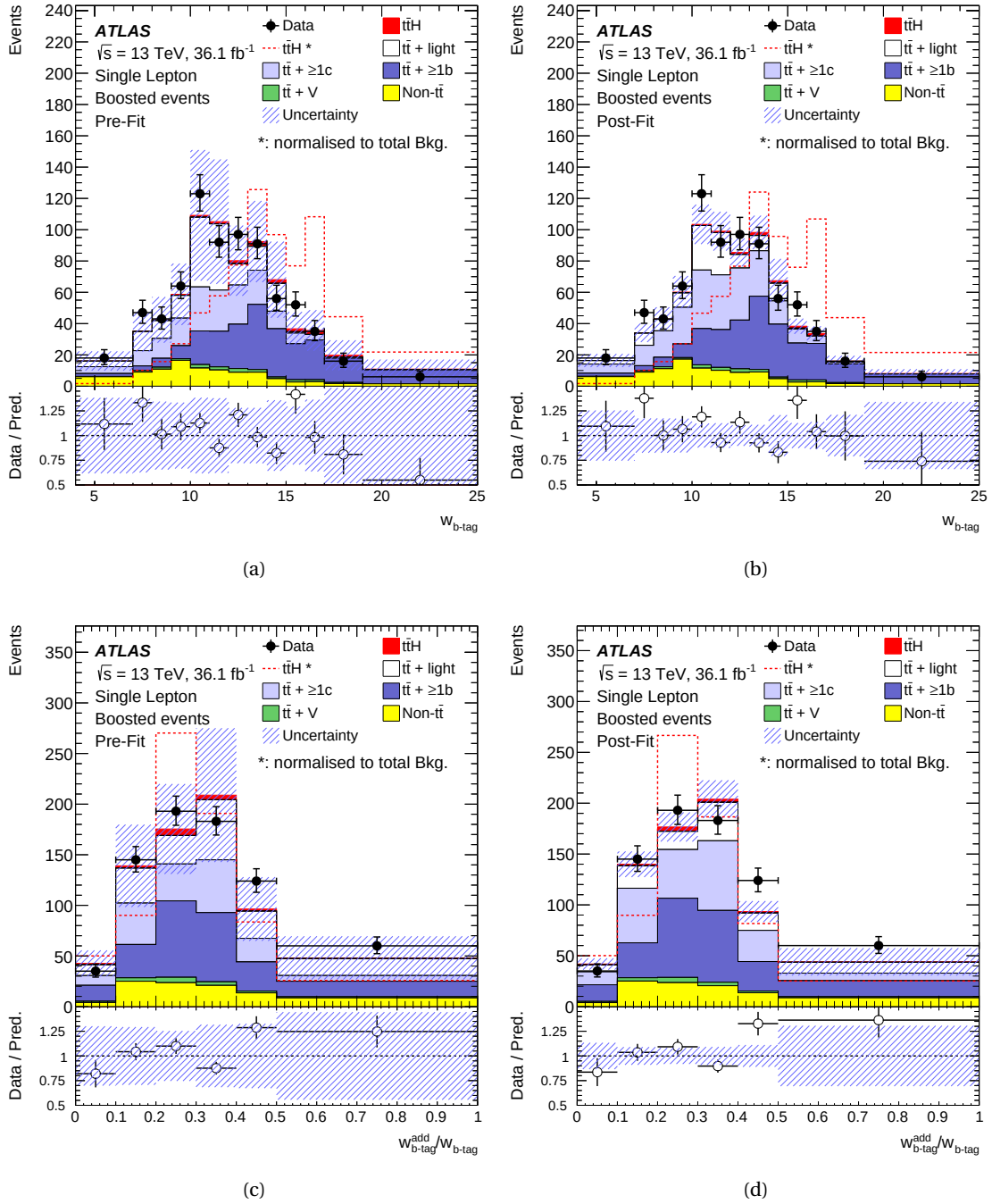


Figure 8.8: The two b-tagging variables used in the boosted classification BDT before (left) and after (right) the full single-lepton fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.

The checks on the agreement between data and prediction before and after running the fit were also carried out for all input variables to the classification BDTs in the resolved signal regions. No significant deviations were found. As expected, the agreement between data and prediction improves post-fit and the uncertainties are reduced.

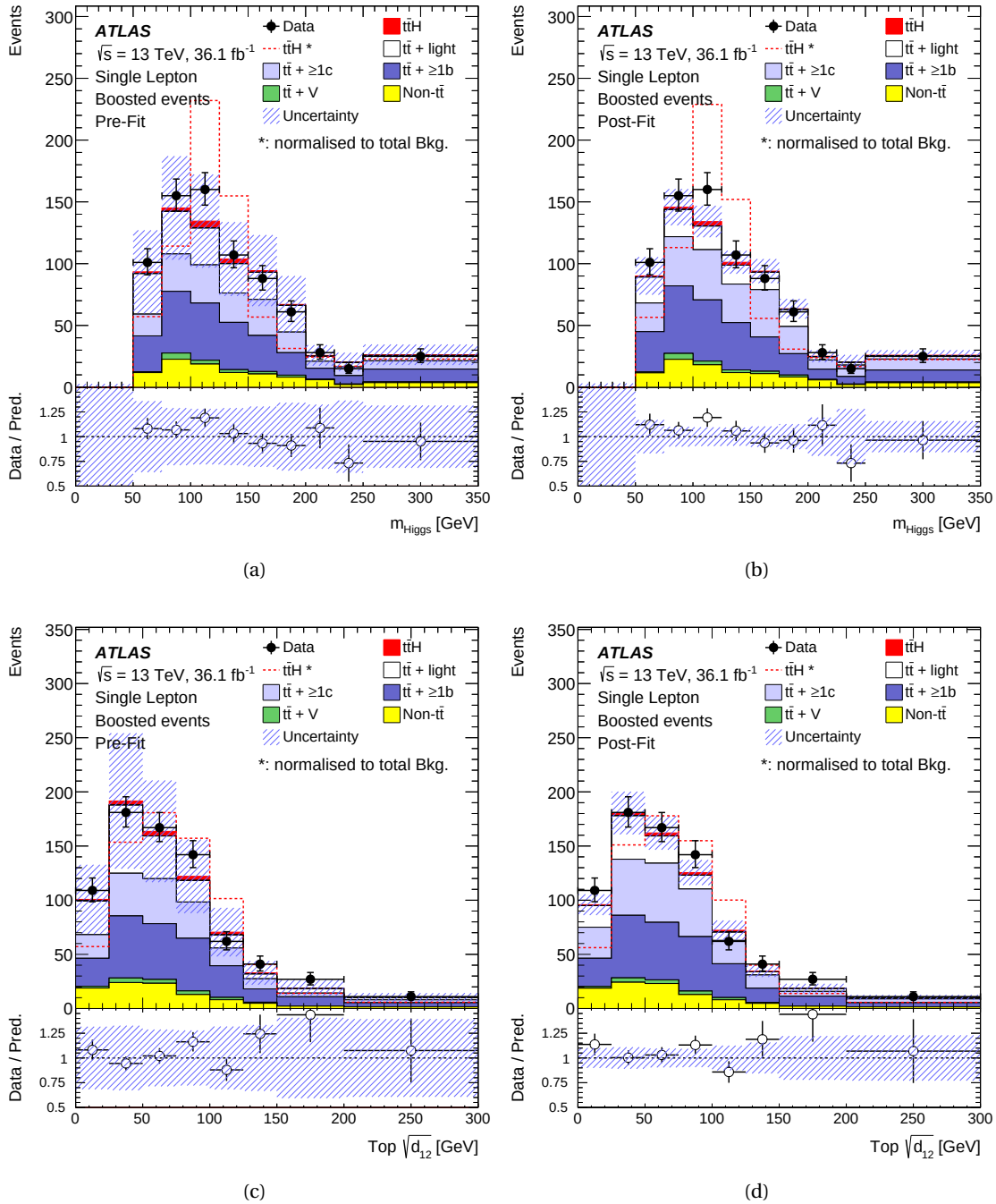


Figure 8.9: The two jet substructure variables used in the boosted classification BDT before (left) and after (right) the full single-lepton fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.

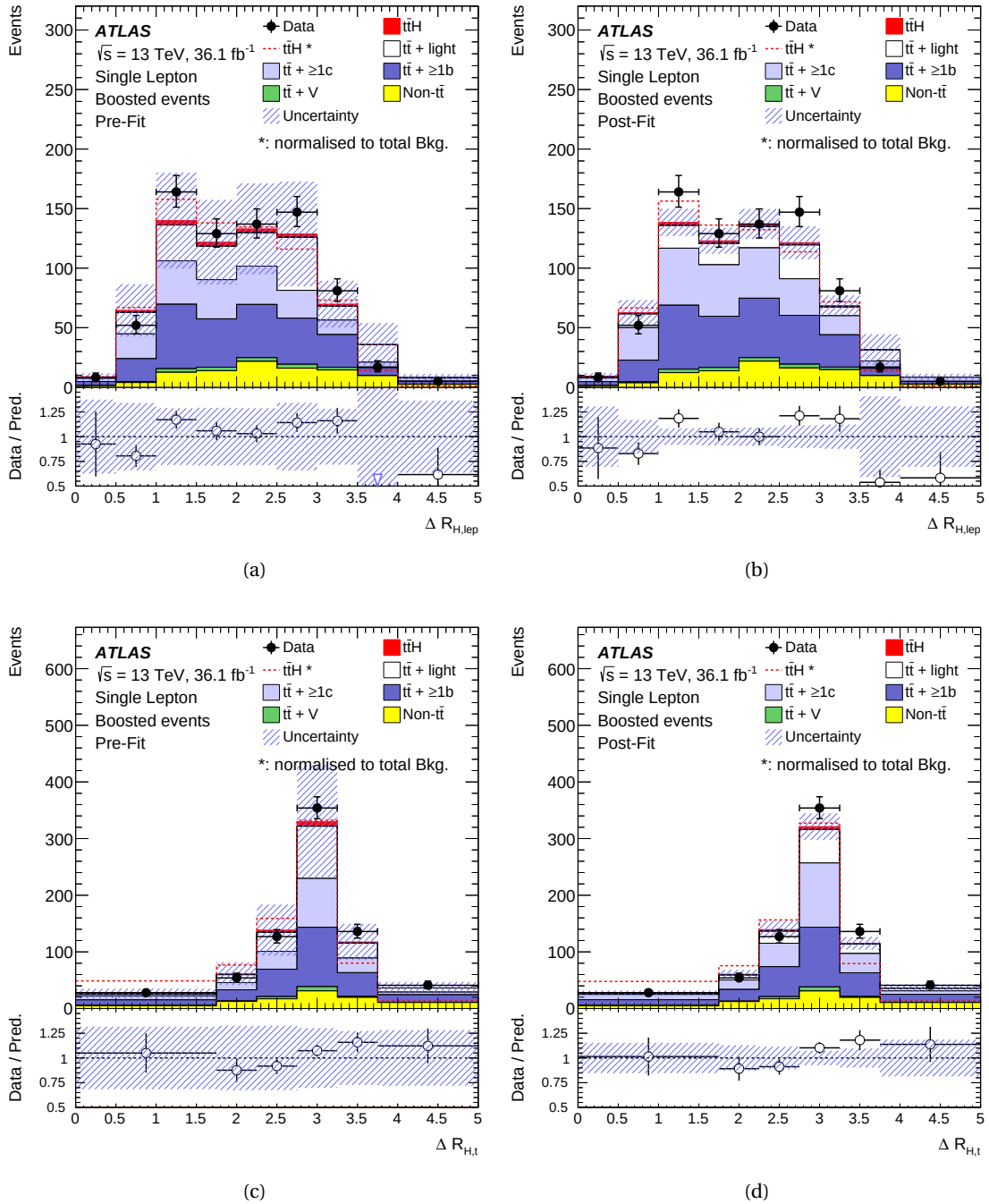


Figure 8.10: Two of the angular variables used in the boosted classification BDT before (left) and after (right) the full single-lepton fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.

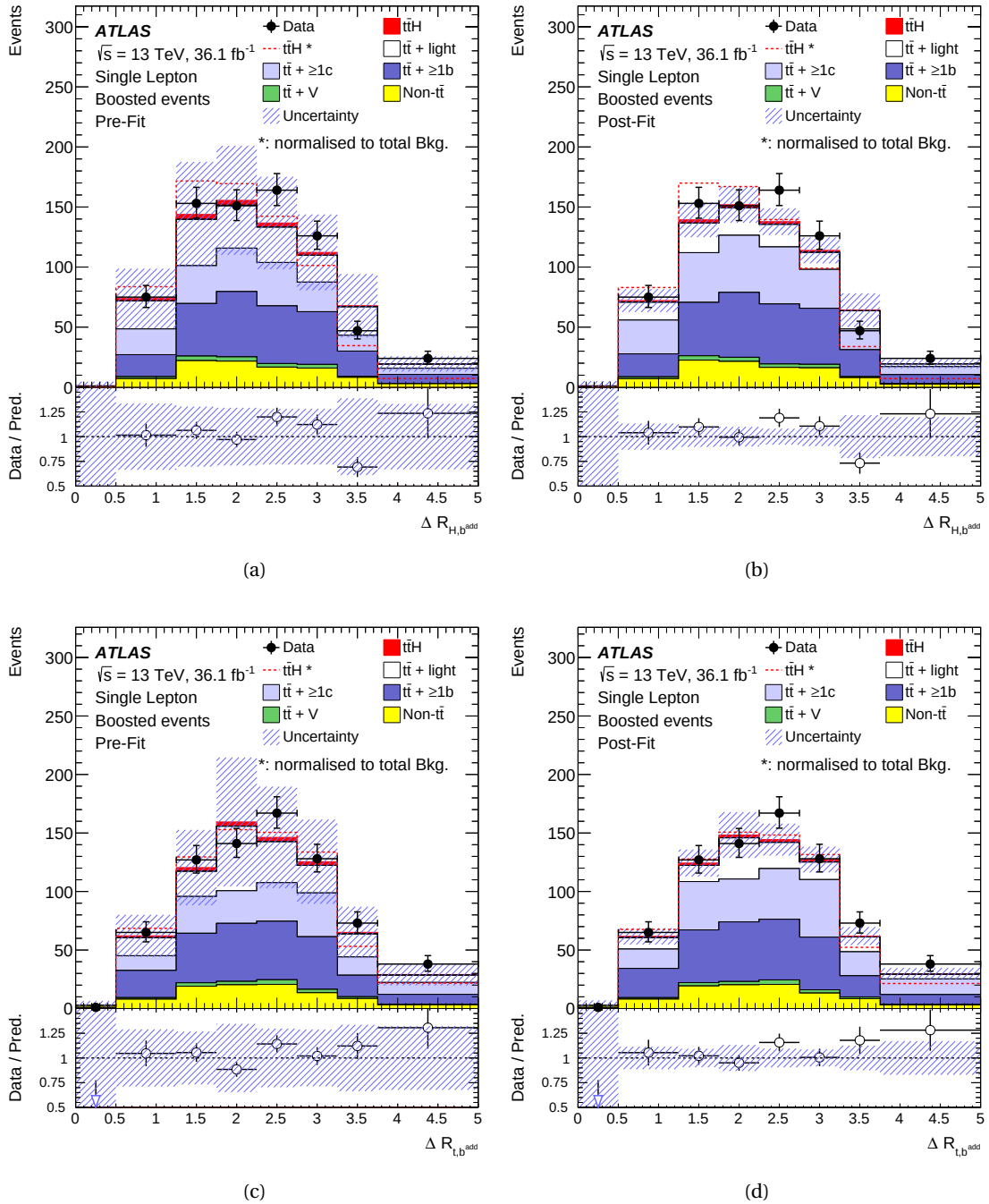


Figure 8.11: Two of the angular variables used in the boosted classification BDT before (left) and after (right) the full single-lepton fit with systematics. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to the fitted μ post-fit. The signal is also shown in a red dashed line where it is normalised to the total background prediction.

The distribution of the final semileptonic boosted classification BDT which enters the fit is shown in figure 8.12 before and after performing the full combined fit to data. The distribution is well modelled by the prediction within the total uncertainty band. Again, we observe an improvement in the modelling agreement as well as a reduction of uncertainties in the post-fit result compared to the pre-fit one. Similar results are observed for the distributions of the resolved classification BDTs as well as the H_T^{had} distributions fitted in two of the resolved control regions (see appendix A.2).

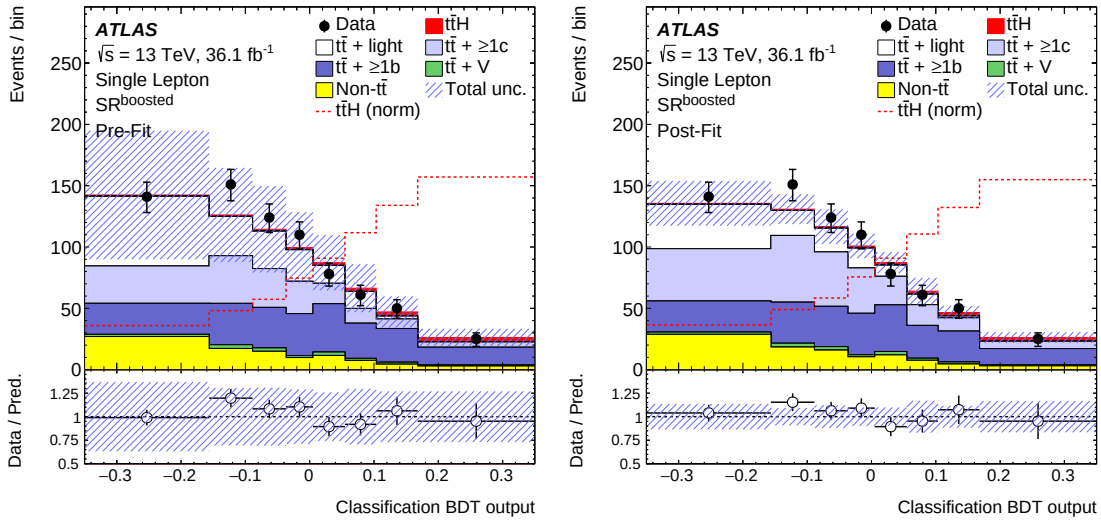


Figure 8.12: Classification BDT output from the semileptonic boosted signal region, showing the distribution before (a) and after (b) the full combined fit to data [103]. The $t\bar{t}H$ signal is shown in red, stacked on top of the background where it is normalised to the SM cross-section pre-fit and to μ post-fit. The signal is shown also in a red dashed line where it is normalised to the total background prediction.

8.2.4 Signal strength and upper limit

The extracted best-fit value for μ from the full combined fit is

$$\mu = 0.84 \pm 0.29(\text{stat.})^{+0.57}_{-0.54}(\text{syst.}) = 0.84^{+0.64}_{-0.61}, \tag{8.3}$$

for a Higgs mass of 125 GeV. The total observed uncertainty on the signal strength is very similar to the expected one from the Asimov fit. An additional fit is performed in which the dilepton and single-lepton channels are decorrelated and assigned two independent signal strength parameters. Note that this two-μ fit preserves the correlations between the two channels for the NPs and normalisation factors. This results in a μ of $0.95^{+0.65}_{-0.62}$ in the single-lepton channel and $-0.24^{+1.02}_{-1.05}$ in the dilepton channel. The negative signal strength measured in the dilepton channel indicates an overestimation of the background by the simulation in regions where signal is expected. This could be due to fluctuations in the simulated sample. When fitting the two channels completely separately, a μ of $0.67^{+0.71}_{-0.69}$ is obtained in the single-lepton channel and a μ of $0.11^{+1.36}_{-1.41}$ in the dilepton channel. These signal strengths are lower than the combined μ

because of the large correlations between the two channels in the systematic uncertainties of the background modelling.

Figure 8.13 shows the signal strength measurements for the semileptonic and dileptonic channel, as well as for the combination of the two. The statistical uncertainty is obtained by fixing all the NPs to their post-fit values (except for μ itself and both of the normalisation factors of $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$) and redoing the fit to data. The total systematic uncertainty is obtained from the total uncertainty by subtracting the statistical uncertainty in quadrature. The systematic uncertainty contributes significantly more to the total uncertainty than the statistical one.

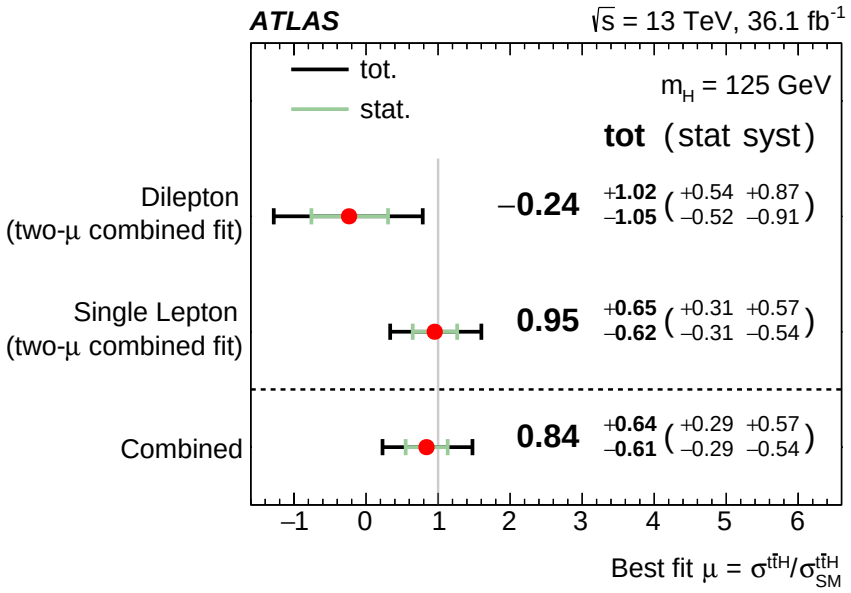


Figure 8.13: Signal strength of μ for the individual semileptonic and dileptonic channels, as well as for the combination [103]. All numbers shown here are taken from the combined fit of the two channels. The numbers from the two channels separately are obtained by keeping the signal strengths uncorrelated and all NPs correlated.

We find an observed significance of 1.4 standard deviations, compared to an expected significance of 1.6σ from the Asimov fit. The CL_s method (see section 6.4) is used to set upper limits on the production of $t\bar{t}H$. A signal strength larger than 2.0 is excluded at the 95% confidence level, as can be seen in figure 8.14. This figure also shows that the combined result is compatible with both the SM hypothesis and the background-only hypothesis within two standard deviations. In addition to the combined upper limit, limits are shown for the decorrelated signal strength parameters of the single-lepton and dilepton channels separately.

8.2.5 Uncertainties

The various contributions to the total uncertainty on μ are listed in table 8.1. The statistical component is obtained by fixing all NPs in the fit to their post-fit value, except for the two normalisation factors and μ itself, and redoing the fit. The intrinsic statistical uncertainty is evaluated by also fixing the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations in this process. The magnitude of

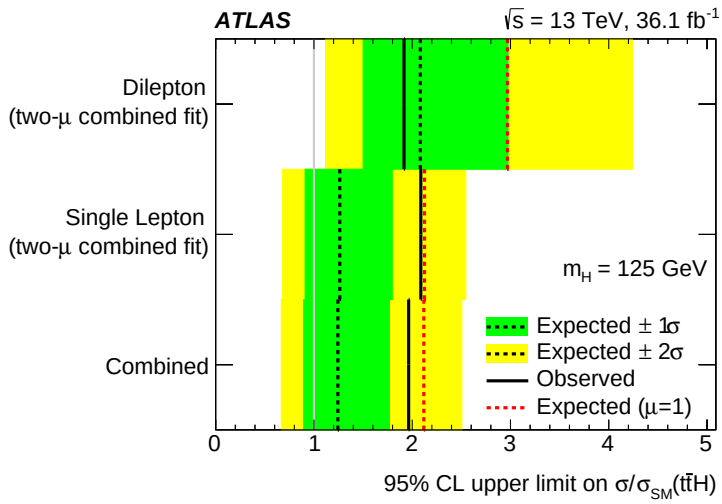


Figure 8.14: Summary of the upper limits on the observed $t\bar{t}H$ cross-section compared to the SM prediction at the 95% confidence level [103]. The limit in the background-only hypothesis is shown by the dotted black line, and the background+signal SM hypothesis is shown by the red dotted line. In the case of the background-only limits, coloured bands are drawn corresponding to the one- and two-standard deviation uncertainty bands.

uncertainties from different sources are obtained by fixing that particular set of uncertainties and redoing the fit. The resulting uncertainty is subtracted in quadrature from the total uncertainty of the full fit in order to determine the component coming from the set of uncertainties under consideration. Note that the total uncertainty is different from the sum in quadrature of all the components because of correlations between NPs generated by the fit. As expected from the Asimov fit, the total uncertainty is dominated by the $t\bar{t}+\geq 1b$ background modelling uncertainties. The second largest source of uncertainty is the statistical uncertainty on the background model. This comes from the limited number of events in the simulated samples and the data-driven estimate of the fakes.

In order to check the flexibility and robustness of the fit, we need to look at the impact, pulls, and constraints of the NPs. Figure 8.15 shows these values for the twenty NPs with the largest impact on μ. The impact of a NP on the sensitivity, $\Delta\mu$, is computed by performing the fit while fixing this NP to its $\pm 1\sigma$ variation, i.e. $\hat{\theta} \pm \Delta\theta$ for pre-fit impact and $\hat{\theta} \pm \Delta\hat{\theta}$ for post-fit impact. The μ computed in this adjusted fit is then compared to the nominal best-fit value of μ to obtain the final impact (see section 6.5).

The first three NPs are the same between the Asimov (figure 8.6) and observed data fit. The NPs' impact on μ is also very similar between the two fits, which gives confidence in the correctness of our fit model. We see again that the uncertainty on the $t\bar{t}H$ signal modelling is non-negligible as it appears at number five in the ranking. There are four uncertainties related to flavour tagging in the top twenty, as well as both of the JER uncertainties. The rest of the systematics in the ranking plot are mostly related to $t\bar{t}$ modelling. In order to check how large the impact of the twenty NPs is on the total uncertainty, a fit was performed where only these twenty sources of systematic uncertainty were included and all others were excluded. The total

Uncertainty source	$\Delta\mu$	
$t\bar{t}+\geq 1b$ modelling	+0.46	−0.46
Background-model stat. unc.	+0.29	−0.31
b-tagging efficiency and mistag rates	+0.16	−0.16
Jet energy scale and resolution	+0.14	−0.14
$t\bar{t}H$ modelling	+0.22	−0.05
$t\bar{t}+\geq 1c$ modelling	+0.09	−0.11
JVT, pile-up modelling	+0.03	−0.05
Other background modelling	+0.08	−0.08
$t\bar{t}+$ light modelling	+0.06	−0.03
Luminosity	+0.03	−0.02
Light lepton (e,μ) id., isolation, trigger	+0.03	−0.04
Total systematic uncertainty	+0.57	−0.54
$t\bar{t}+\geq 1b$ normalisation	+0.09	−0.10
$t\bar{t}+\geq 1c$ normalisation	+0.02	−0.03
Intrinsic statistical uncertainty	+0.21	−0.20
Total statistical uncertainty	+0.29	−0.29
Total uncertainty	+0.64	−0.61

Table 8.1: The various contributions to the uncertainty on μ. The background-model statistical uncertainty refers to the statistical uncertainty in the simulated events and data-driven estimate of the fakes in the single-lepton channel. The intrinsic statistical uncertainty refers to the statistical uncertainty evaluated after fixing all NPs in the fit including the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisations.

uncertainty on μ decreases by 5% in this case.

Figure 8.15 also shows the best-fit value and post-fit uncertainty for the twenty NPs with the largest impact on μ. Unlike in the Asimov fit, we see that most NPs are pulled away from their nominal value in the fit to data. However, no systematic is pulled by more than $\Delta\theta$ from its nominal value. In order to understand these shifts, fits are performed in which the shifted NPs are one-by-one set to be uncorrelated between analysis regions and between simulated event samples. The pulls were found to mainly correct the simulated $t\bar{t}$ background to the observed data. None of the regions was found to be the main contributor to any of the observed pulls. Similar pulls are observed in a background-only fit in which the most signal rich bins are removed from the discriminant distributions. The variations in μ due to the pulls are quantified by fixing the corresponding NPs to their pre-fit values, redoing the fit, and comparing the obtained μ to the nominal one. These variations were found to be smaller than the total uncertainty on the signal strength μ.

We also see that some NPs are significantly constrained by the fit, i.e. their post-fit uncertainty is small compared to their pre-fit uncertainty. This happens mainly in the NPs associated with the modelling of $t\bar{t}$+HF ($t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$). A NP is constrained by the fit when the associated uncertainty affects the discriminant distributions in such a way that large deviations from data arise. The capability of the fit to constrain the systematics was validated in the fits to Asimov data and pseudo data, as described above. No additional or larger constraints are seen when fitting on observed data compared to Asimov data.

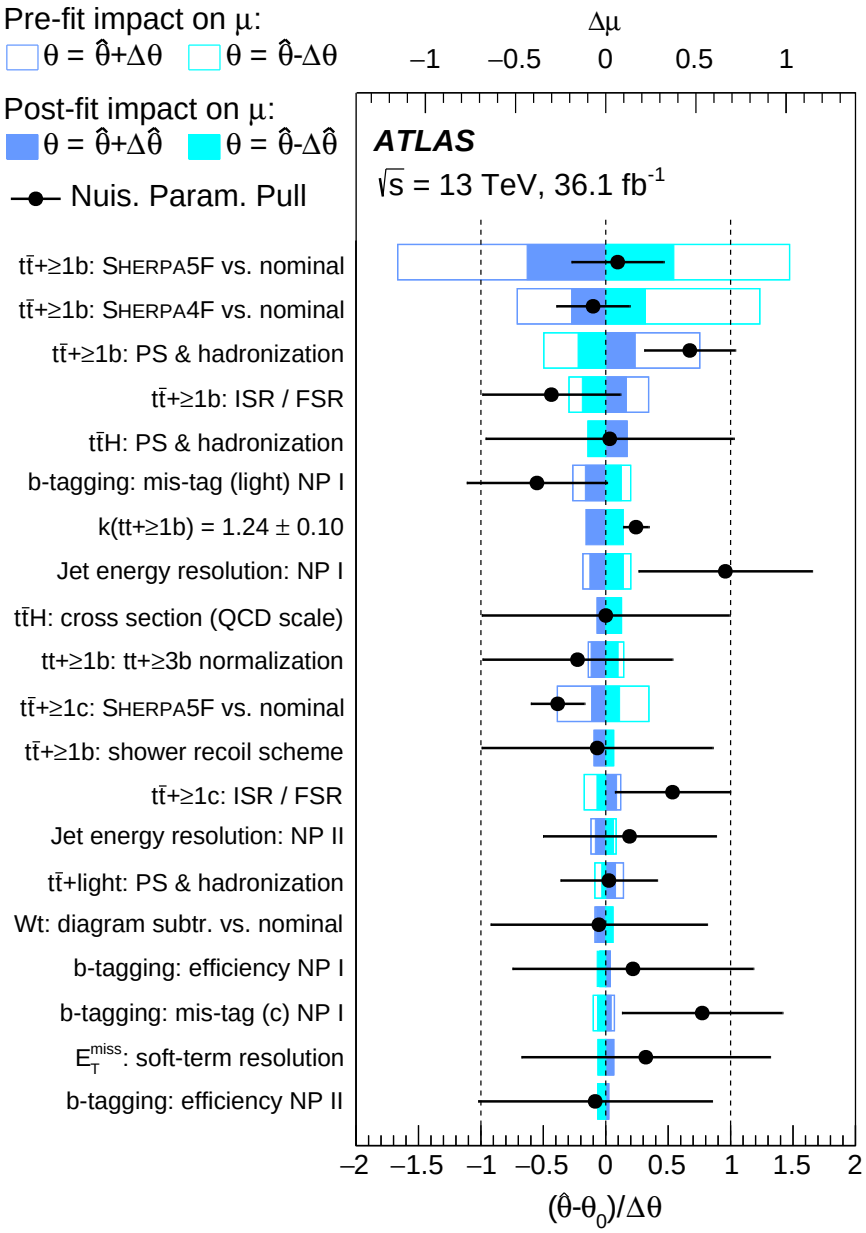


Figure 8.15: The top 20 nuisance parameters in the fit, ranked according to their impact ($\Delta\mu$) on the measured signal strength μ [103]. Both the pre-fit (empty blue rectangles) and post-fit (filled blue rectangles) impact on μ are shown and correspond to the top axis. The black dots denote the pulls of the nuisance parameters from their nominal value θ_0 with the black bar indicating the post-fit error relative to the pre-fit error, after applying the constraints from the fit. The pulls and constraints correspond to the bottom axis.

8.3 Combination with other $t\bar{t}H$ searches in ATLAS

The $t\bar{t}H(H \to b\bar{b})$ analysis was combined with three other $t\bar{t}H$ production searches to achieve the best possible significance. These other searches are tailored to different Higgs decay modes:

- $t\bar{t}H(H \to \text{ML})$ in which ML refers to *multilepton*. This channel uses seven final states distinguished by the number and flavour of charged lepton candidates [105].
- $t\bar{t}H(H \to \gamma\gamma)$ in the semileptonic, dileptonic and all-hadronic $t\bar{t}$ channels [126].
- $t\bar{t}H(H \to ZZ^* \to 4l)$ in a single category including all $t\bar{t}$ channels [127].

The overlap between the analyses is negligible due to carefully selected preselection cuts. The combination is constructed as a product of the likelihood functions from the four individual analyses, based on simultaneous fits to the signal and control regions of each analysis.

8.3.1 Evidence for $t\bar{t}H$

The combination is first carried out with each analysis using the same ATLAS dataset of 36.1 fb^{-1} from 2015–2016 [105]. All analyses use the same theoretical prediction, MC event generator, and associated uncertainties of the signal $t\bar{t}H$ process. The $H \to b\bar{b}$ and $H \to$ ML analyses select a negligible amount of events of other Higgs boson production mechanisms. However, the $H \to \gamma\gamma$ and $H \to ZZ$ analyses measure the $t\bar{t}H$ signal strength in a global analysis of all Higgs boson production modes. In the combined $t\bar{t}H$ result, only the $t\bar{t}H$-enriched categories from these latter two analyses are included. These $t\bar{t}H$-enriched regions suffer from significant contamination of other Higgs boson production modes. For the extraction of the $t\bar{t}H$ signal strength, all non-$t\bar{t}H$ production modes are considered as background and their cross-sections, together with all Higgs boson decay branching fractions, are set to the SM expectations with appropriate theoretical uncertainties [14]. This includes the single top Higgs boson production processes *tHjb* and *WtH*. This method results in slightly different μ values reported in the combination compared to the ones reported in the individual analyses.

The majority of NPs associated with the same systematic uncertainty sources are treated as correlated between the four analyses. None of the NPs in the fit are strongly constrained by more than one analysis, and the value of the signal strength obtained from the combined fit does not depend on the choice of the correlation scheme. The correlations are treated as follows:

- **Experimental uncertainties:** The systematic uncertainties related to the JES are correlated between all analyses except for the NP associated with the fraction of jets initiated by quarks and gluons. This fraction is significantly different between the four analyses due to the different event selections. NPs related to the JER are correlated between all analyses except for the control regions of the $H \to b\bar{b}$ analysis. As described in section 7.3, the JER uncertainty in the $H \to b\bar{b}$ analysis is divided into two independent components; one for the semileptonic $\text{CR}^{5j}_{t\bar{t}+\geq 1c}$ and dileptonic $\text{CR}^{3j}_{t\bar{t}+\text{light}}$ regions, and one for all the other regions. This is done because the JER uncertainties in the first two regions show different

behaviour from the other regions in the fit. Decorrelating this uncertainty in these control regions therefore avoids constraining this systematic in the signal regions and gives a conservative estimate of its impact. The uncertainties related to flavour-tagging are correlated between the $H \to b\bar{b}$ and $H \to$ ML analyses and between the $H \to \gamma\gamma$ and $H \to ZZ$ analyses. This is because the first pair uses a different calibration scheme for the flavour-tagging efficiencies compared to the latter pair. All other experimental uncertainties are treated as correlated between the four analyses.

- **Theoretical uncertainties:** All theoretical uncertainties associated with the Higgs boson production cross-sections and decay branching fractions are fully correlated between the analyses. The uncertainties on the cross-sections and modelling of the backgrounds in the $H \to b\bar{b}$ and $H \to$ ML searches are correlated between the two analyses. However, the additional modelling uncertainties on the $t\bar{t}$+HF backgrounds from the $H \to b\bar{b}$ search are not applied to any of the other analyses because the phase space of the other searches are not as sensitive to the modelling of this background. The three other channels have their own independent set of systematic uncertainties for the modelling of this background.

The final observed and expected μ measurements, as well as the significance for $t\bar{t}H$ production, are shown in table 8.2. The observed best-fit value for μ from the combined likelihood function is

$$\mu_{t\bar{t}H} = 1.17 \pm 0.19(\text{stat.})^{+0.27}_{-0.23}(\text{syst.}) = 1.17^{+0.33}_{-0.30}. \tag{8.4}$$

The background-only hypothesis is excluded at 4.2σ, with an expectation of 3.8σ when assuming the SM $t\bar{t}H$ prediction. This result constitutes evidence for $t\bar{t}H$ production.

Channel	Best-fit μ		Significance	
	Observed	Expected	Observed	Expected
$H \to$ ML	$1.6^{+0.5}_{-0.4}$	$1.0^{+0.4}_{-0.4}$	4.1σ	2.8σ
$H \to b\bar{b}$	$0.8^{+0.6}_{-0.6}$	$1.0^{+0.6}_{-0.6}$	1.4σ	1.6σ
$H \to \gamma\gamma$	$0.6^{+0.7}_{-0.6}$	$1.0^{+0.8}_{-0.6}$	0.9σ	1.7σ
$H \to 4l$	< 1.9	$1.0^{+3.2}_{-1.0}$	-	0.6σ
Combined	$1.2^{+0.3}_{-0.3}$	$1.0^{+0.3}_{-0.3}$	4.2σ	3.8σ

Table 8.2: Observed and expected best-fit μ measurements and the $t\bar{t}H$ production significance from the four $t\bar{t}H$ analyses combined for this search. The $t\bar{t}H(H \to ZZ^* \to 4l)$ analysis observed no events, so a 68% confidence level upper limit on μ is reported.

The observed best-fit value of the signal strength corresponds to a $t\bar{t}H$ cross-section of

$$\sigma_{t\bar{t}H} = 590^{+160}_{-150}\ \text{fb}, \tag{8.5}$$

which is compatible with the SM prediction of 507^{+35}_{-50} fb [14]. Figure 8.16 shows a breakdown of the extracted μ values for each of the four analyses, plus the combined result. Since the

$t\bar{t}H(H \to ZZ^* \to 4l)$ analysis observed no events, a 68% (1σ) confidence level upper limit on μ is reported for this decay mode, computed with the CL_s method.

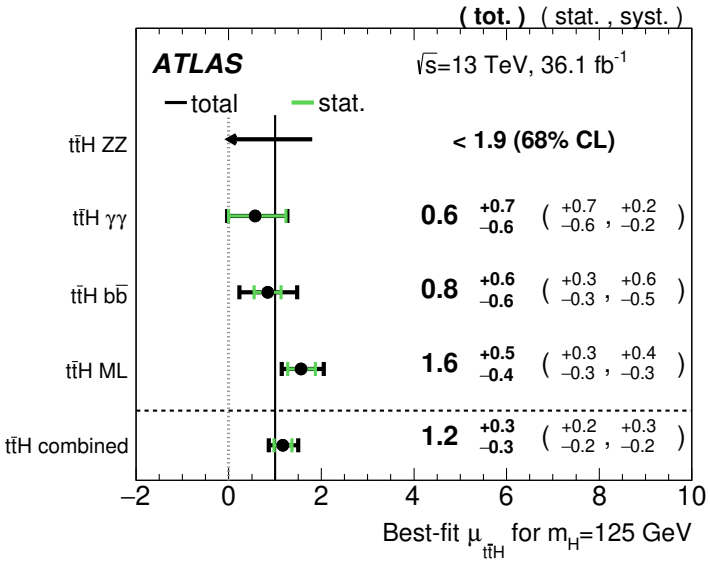


Figure 8.16: Summary of the observed best-fit μ measurements from the four $t\bar{t}H$ analyses combined for this search [105]. The total uncertainty is broken down into its statistical and systematic components. The $t\bar{t}H(H \to ZZ^* \to 4l)$ analysis observed no events, so a 68% confidence level upper limit on μ is reported. The black vertical line indicates the SM expectation.

The impact of the various uncertainties on the combined signal strength are shown in table 8.3. The combined $t\bar{t}H$ search is dominated by the systematic component of the uncertainty. The overall dominant systematic uncertainty is the $t\bar{t}$ modelling in the $H \to b\bar{b}$ analysis. The uncertainty on the $t\bar{t}H$ signal modelling and cross-section has the second-largest impact on the final result.

8.3.2 Observation of $t\bar{t}H$

After the establishment of evidence for the $t\bar{t}H$ process from the combination of the four decay channels using 36 fb^{-1} of data, a second combination is carried out in which the $H \to \gamma\gamma$ and $H \to ZZ$ analyses are updated with data collected in 2017 [128]. Additionally, a combination is performed with results based on 4.5 fb^{-1} of $\sqrt{s}$ = 7 TeV data from 2011 and 20.3 fb^{-1} of $\sqrt{s}$ = 8 TeV data from 2012 [129]. For the latter combination, the SM expectations for the cross-sections and branching ratios are updated with the values in reference [14].

The $H \to \gamma\gamma$ analysis uses a dataset of 79.8 fb^{-1} at $\sqrt{s}$ = 13 TeV. The sensitivity is greatly enhanced compared to that in reference [126] due to an improved BDT performance, better lepton and photon reconstruction algorithms [130], and a re-evaluated event selection and categorisation. The $H \to ZZ$ search uses the same dataset as the $H \to \gamma\gamma$ analysis. It improves upon its result quoted in reference [127] by using better lepton and photon reconstruction algorithms [130], and by defining two $t\bar{t}H$-enriched signal regions and applying a BDT in one of them.

Uncertainty source	**$\Delta\mu$**	
$t\bar{t}$ modelling in $H \to b\bar{b}$ analysis	+0.15	-0.14
$t\bar{t}H$ modelling (cross-section)	+0.13	-0.06
Non-prompt light-lepton and fake τ_{had} estimates	+0.09	-0.09
Simulation statistics	+0.08	-0.08
Jet energy scale and resolution	+0.08	-0.07
$t\bar{t}V$ modelling	+0.07	-0.07
$t\bar{t}H$ modelling (acceptance)	+0.07	-0.04
Other non-Higgs boson backgrounds	+0.06	-0.05
Other experimental uncertainties	+0.05	-0.05
Luminosity	+0.05	-0.04
Jet flavour tagging	+0.03	-0.02
Modelling of other Higgs boson production modes	+0.01	-0.01
Total systematic uncertainty	+0.27	-0.23
Statistical uncertainty	+0.19	-0.19
Total uncertainty	+0.34	-0.30

Table 8.3: Summary of the uncertainties affecting the value of μ from the combined likelihood fit across all four $t\bar{t}H$ channels.

The correlation scheme for all systematic uncertainties between the $H \to b\bar{b}$ and $H \to$ ML analyses are kept the same as described above for the $t\bar{t}H$ evidence combination (see reference [105]). This is also the case for all theoretical systematic uncertainties and their correlations between the four searches. On the other hand, the experimental systematics are evaluated individually for most sources because the updated analyses use new reconstruction software compared to the $H \to b\bar{b}$ and $H \to$ ML analyses. Some components of the experimental uncertainties are correlated between the channels. Just as in the 36.1 fb^{-1} combination, all non-$t\bar{t}H$ production modes are considered as background and their cross-sections are set to the SM expectations with appropriate theoretical uncertainties. All decay branching fractions of the Higgs boson are set to their SM expectations as well.

The results on the signal strength from the full combination of the four searches with 13 TeV data are shown in figure 8.17. The full combined signal strength from the likelihood fit across the four channels has a best-fit value of

$$\mu_{t\bar{t}H} = 1.32 \pm 0.18(\text{stat.})^{+0.21}_{-0.19}(\text{syst.}) = 1.32^{+0.28}_{-0.26}. \tag{8.6}$$

This signal strength corresponds to an observed (expected) excess of $t\bar{t}H$ relative to the background-only hypothesis of 5.8σ (4.9σ).

The total uncertainty on the final result is dominated by its systematic component. Just as in the combination carried out in reference [105] described above, the uncertainties with the largest impact on the result arise from the modelling of $t\bar{t}$+HF in the $H \to b\bar{b}$ analysis and the modelling of the signal $t\bar{t}H$ process.

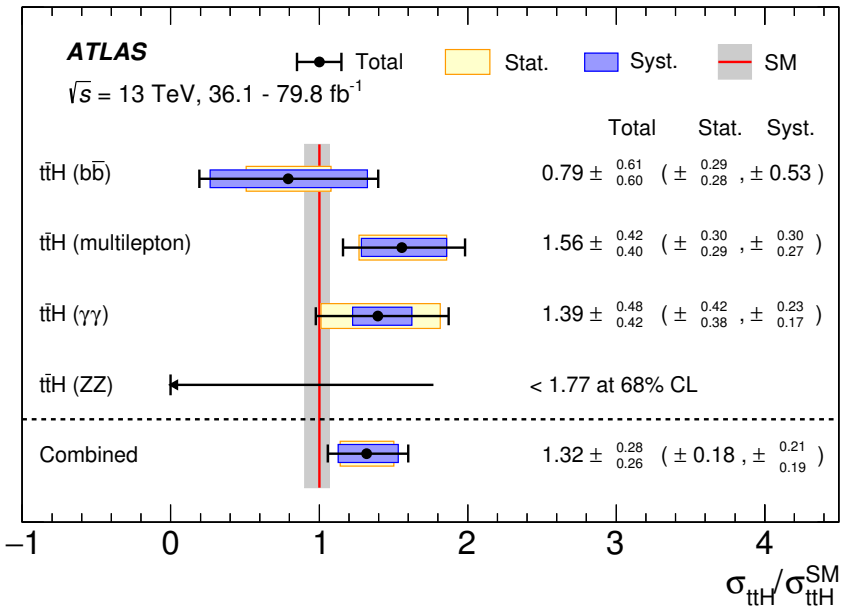


Figure 8.17: Summary of the observed best-fit μ measurements from the four $t\bar{t}H$ analyses combined for this search, using 13 TeV data [128]. The black lines show the total uncertainty, which is broken down into its statistical and systematic components. The $t\bar{t}H(H \to ZZ^* \to 4l)$ analysis observed no events, so a 68% confidence level upper limit on μ is reported. The grey band around the red line indicating the SM prediction represents the uncertainties on the PDF, α_s, and missing higher-order corrections.

The measured $t\bar{t}H$ cross-sections of all four channels and the combination at 13 TeV are listed in table 8.4. The observed and expected significances are also listed for each of these, as well as for the combination using 7, 8, and 13 TeV data. The likelihood fit across all channels using 7, 8, and 13 TeV data results in an observed (expected) excess of $t\bar{t}H$ relative to the background-only hypothesis of 6.3σ (5.1σ). This constitutes direct observation of the production of the Higgs boson in association with a top quark pair.

Channel	**Integrated luminosity [fb^{-1}]**	**$t\bar{t}H$ cross-section [fb]**	**Significance**	
			Obs.	**Exp.**
$H \to$ ML	36.1	$790 \pm 150(\text{stat.})^{+150}_{-140}(\text{syst.})$	4.1σ	2.8σ
$H \to b\bar{b}$	36.1	$400^{+150}_{-140}(\text{stat.}) \pm 270(\text{syst.})$	1.4σ	1.6σ
$H \to \gamma\gamma$	79.8	$710^{+210}_{-190}(\text{stat.})^{+120}_{-90}(\text{syst.})$	4.1σ	3.7σ
$H \to ZZ$	79.8	< 900(68%CL)	-	1.2σ
Combined (13 TeV)	36.1-79.8	$670 \pm 90(\text{stat.})^{+110}_{-100}(\text{syst.})$	5.8σ	4.9σ
Combined (7, 8, 13 TeV)	4.5, 20.3, 36.1-79.8	-	6.3σ	5.1σ

Table 8.4: Measured total $t\bar{t}H$ cross-sections at $\sqrt{s} = 13$ TeV alongside the observed and expected significance. The results include the four individual searches, the 13 TeV combination, and the significance of the combination with 7, 8, and 13 TeV data. The $t\bar{t}H(H \to ZZ^* \to 4l)$ analysis observed no events, so a 68% confidence level upper limit on the $t\bar{t}H$ cross-section is reported.

The cross-sections measured at centre-of-mass energies of 8 TeV and 13 TeV are compared to the SM prediction in figure 8.18. The measured cross-section at 8 TeV is $220 \pm 100(\text{stat.}) \pm 70(\text{syst.})$ fb. The combination of 13 TeV data measures a total cross-section of $670 \pm 90(\text{stat.})^{+110}_{-100}(\text{syst.})$ fb which is in agreement with the SM prediction of 507^{+35}_{-50} fb [14]. This measurement establishes a direct observation of the Yukawa coupling between the Higgs boson and the top quark.

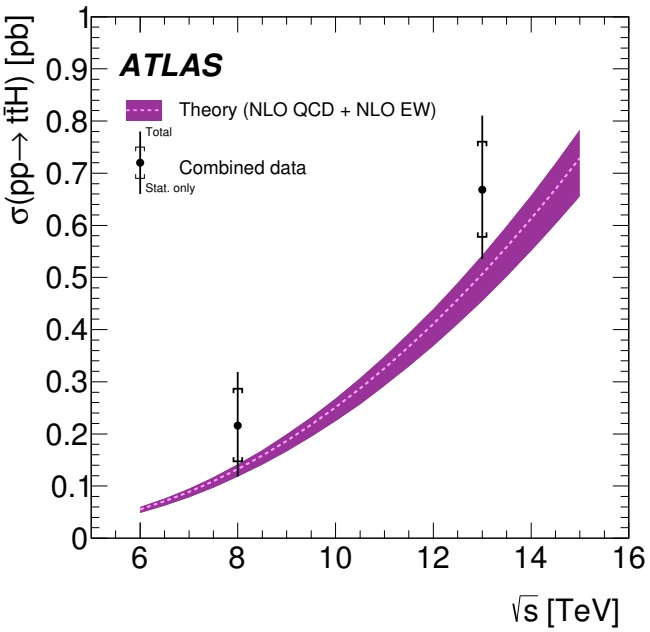


Figure 8.18: The measured $t\bar{t}H$ cross-section at $\sqrt{s} = 8$ TeV and 13 TeV [128]. The black vertical lines show the total uncertainty and its statistical component. The purple band around the SM theory prediction represents the uncertainties on the PDF, α_s, and missing higher-order corrections.

9 Conclusions

The last piece of the Standard Model puzzle was the observation of the Higgs boson in 2012. To date, all measurements of this particle's properties have been consistent with the Standard Model predictions. One very interesting place to look for new physics is in the coupling of the Higgs boson to the top quark. Since the top quark is the heaviest fermion, this is the largest Yukawa coupling in the model and is predicted to be close to one. The production of a Higgs boson in association with a top quark pair gives direct access to this coupling. The $t\bar{t}H$ process was not found in Run I of the LHC, when the centre-of-mass energy was 7 – 8 TeV. This thesis presented a search for the $t\bar{t}H$ process in Run II, in which the energy was increased to 13 TeV. This rise in energy leads to a fourfold increase of the predicted $t\bar{t}H$ cross-section and thus opens up the playing field for investigations of the top Yukawa coupling.

The search presented here was designed for the Higgs decaying to a pair of bottom quarks, but all Higgs decay channels were treated as signal. The $H \to b\bar{b}$ decay mode was chosen because it has the largest branching ratio in the Standard Model of 58%. The analysis was divided into the semileptonic and dileptonic decays of the $t\bar{t}$ system. The main focus of this thesis was the semileptonic boosted channel, in which the Higgs boson and the hadronically decaying top quark are produced at high transverse momentum compared to their mass. This channel was added to the $t\bar{t}H(H \to b\bar{b})$ analysis for the first time and significantly reduces the combinatorics of jets in the final state of the $t\bar{t}H$ process. This makes it easier to reconstruct the objects in the events. With a simple jet selection in the boosted channel, the correct Higgs candidate was found in 47% of signal events. This is comparable to the resolved analysis which used a complicated reconstruction BDT and found the correct Higgs candidate 48% of the times.

The $t\bar{t}H$ process constitutes only 1% of the total Higgs production mechanisms and suffers from large backgrounds. The main background comes from top quark pairs produced with additional jets. The resulting low signal-to-background ratio necessitated the use of multivariate techniques to separate signal from background events. In the boosted region, a BDT was used to this end with eight discriminating variables including the Higgs candidate mass.

The final results were obtained from a profile likelihood fit across all semilepton and dilepton regions, using a dataset of 36.1 fb^{-1} taken with the ATLAS detector in 2015 and 2016. The best-

fit value of the signal strength was found to be

$$\mu = 0.84 \pm 0.29(\text{stat.})^{+0.57}_{-0.54}(\text{syst.}) = 0.84^{+0.64}_{-0.61},$$

for a Higgs mass of 125 GeV. This corresponds to a significance of 1.4 standard deviations. A signal strength larger than 2.0 was excluded at the 95% confidence level.

The results from the $b\bar{b}$ decay channel were combined with three other $t\bar{t}H$ searches optimised for the multilepton, $\gamma\gamma$, and ZZ decay modes. The combination in which all analyses used 36 fb^{-1} of data led to a significance of 4.2σ above the background-only hypothesis. This constitutes evidence for $t\bar{t}H$ production and corresponds to a cross-section of

$$\sigma_{t\bar{t}H} = 590^{+160}_{-150} \text{ fb},$$

which is compatible with the SM prediction of 507^{+35}_{-50} fb.

The combination was repeated with the $\gamma\gamma$ and ZZ decay channels updated to include ATLAS data from 2017, leading to a total luminosity of 79.8 fb^{-1} for these two analyses. This resulted in an observed (expected) excess of 5.8σ (4.9σ). The best-fit value for the signal strength parameter found was

$$\mu = 1.32 \pm 0.18(\text{stat.})^{+0.21}_{-0.19}(\text{syst.}) = 1.32^{+0.28}_{-0.26},$$

corresponding to a cross-section of $\sigma_{t\bar{t}H} = 670^{+142}_{-135}$ fb. The Run II results were combined with the Run I dataset to achieve an observed (expected) significance of 6.3σ (5.1σ). This marked the first direct observation of the Higgs coupling to the top quark.

The full $t\bar{t}H(H \to b\bar{b})$ analysis was shown to be limited by systematics, the most notorious of which was the modelling of the $t\bar{t}+ \geq 1b$ background process. A better understanding of this process will be needed to enhance the sensitivity of this analysis in the future. Precise measurements of this process at 13 TeV could be used as input to MC generators in order to improve their modelling. Another useful step would be to merge the NLO calculation of $t\bar{t}+ \geq 1b$ from SHERPA+OPENLOOPS with the inclusive $t\bar{t}$+jets sample in order to profit fully from this very precise calculation. The second largest source of uncertainty in the analysis was the statistical uncertainty on the simulated background samples. This can be improved in the future by generating more MC events, especially in the small region of phase space where the signal is present.

The boosted region did not add significant sensitivity to the $t\bar{t}H(H \to b\bar{b})$ analysis. However, a larger dataset at the end of Run II and beyond will allow for a tighter signal region definition to increase its purity. The use of jet tagging techniques in the definition of the signal region, as well as more advanced jet substructure variables in the BDT, could also greatly improve the sensitivity of this region. The boosted regime will become more important as the LHC continues to run and can in the future be used for a differential cross-section measurement in the high-p_T phase space region.

APPENDIX A

A.1 Dilepton event selection details

The dilepton $t\bar{t}H(H \to b\bar{b})$ analysis has three signal regions and four control regions in total. The signal regions all fall in the ≥ 4 jets category (see figure A.1(b)). The purest of these is the SR_1 and has three jets passing the *very tight* b-tagging WP and a fourth jet passing either the *very tight* or *tight* selection. Two of the control regions are in the ≥ 4 jets category (enriched in $t\bar{t}+\geq 1c$ and $t\bar{t}+$ light, see figure A.1(b)) and two in the 3 jets category (enriched in $t\bar{t}+\geq 1b$ and $t\bar{t}+$ light, see figure A.1(a)). The background composition of the dilepton signal and control regions is shown in figure A.2 and the purity of each region is shown in figure A.3.

A.2 Pre-fit and post-fit distributions used in the combined fit

The final results of the $t\bar{t}H(H \to b\bar{b})$ analysis are obtained from a profile likelihood fit over the nine signal regions and ten control regions of the semileptonic and dileptonic channels combined. In each of the signal regions, a dedicated classification BDT is used. The BDT from the boosted signal region was shown in figure 8.12. Figures A.4, A.5, and A.6 show the classification BDT distributions for all the resolved signal regions, both before and after the full combined fit to data.

The H_T^{had} distribution (the scalar sum of the jet p_T) is fitted in two of the resolved semileptonic control regions enriched in $t\bar{t}+\geq 1c$. In all other resolved control regions, only one bin is fitted which represents the total event yield. This choice was made because the H_T^{had} variable is not well-modelled in these control regions and the mismodelling is not fully covered by the uncertainties. The distributions for the H_T^{had} variables in the single-lepton $t\bar{t}+\geq 1c$ enriched control regions are shown in figure A.7 before and after performing the full combined fit to data.

All the distributions in the signal and control regions are well modelled before the fit and have an improved agreement between data and prediction after the fit due to the NPs being adjusted. The post-fit uncertainty is reduced due to the constraints on and correlations between the NPs generated by the fit.

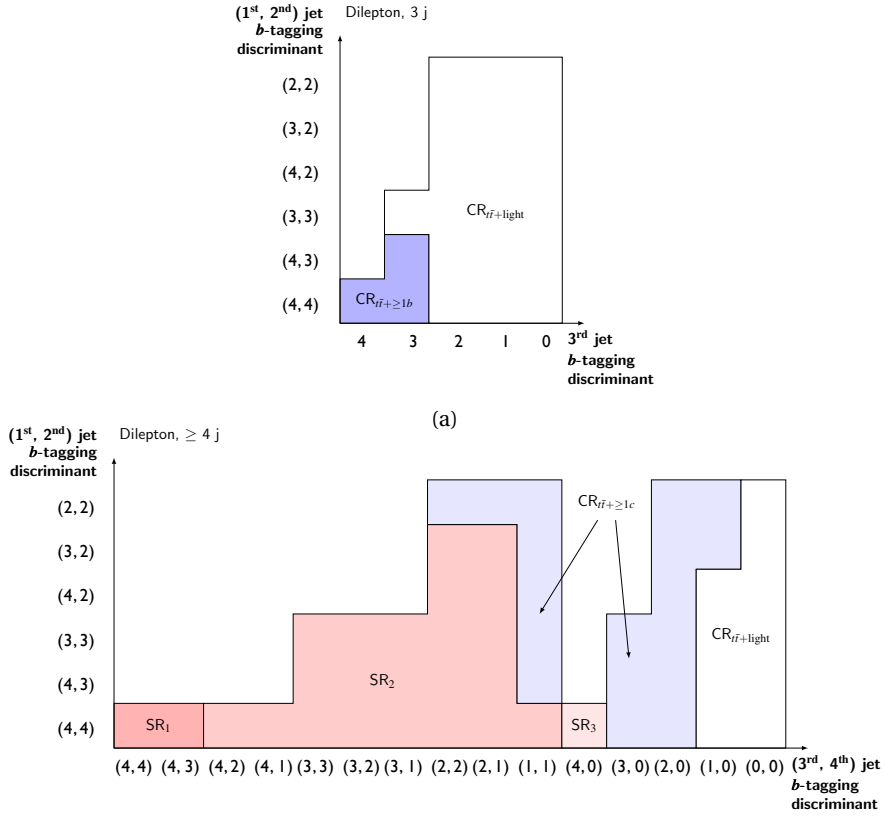


Figure A.1: Definition of the regions for the dileptonic resolved channel, for events containing exactly 3 jets (a) and events containing at least 4 jets (b) [103]. The vertical axis shows the b-tagging discriminant value for the first and second jets whereas the horizontal axis shows this discriminant for the third (and fourth) jets. The jet ordering is based on the value of this discriminant in descending order.

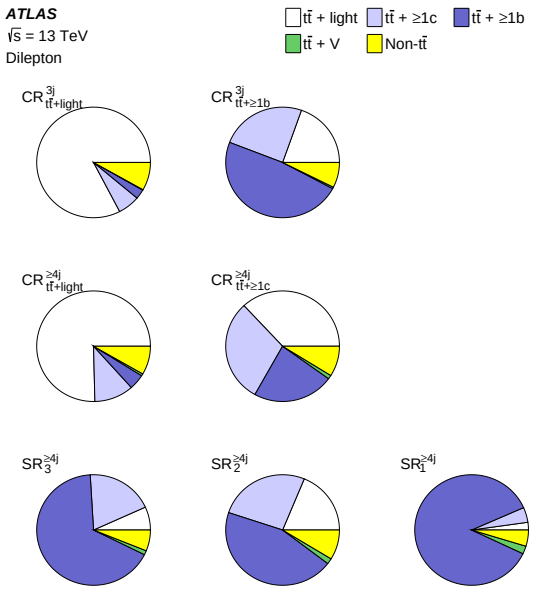


Figure A.2: Background composition of the dileptonic signal and control regions [103]. The $t\bar{t}$ background is divided into $t\bar{t}$+ light, $t\bar{t}+\geq 1c$, $t\bar{t}+\geq 1b$, and $t\bar{t}+\mathrm{V}$ contributions.

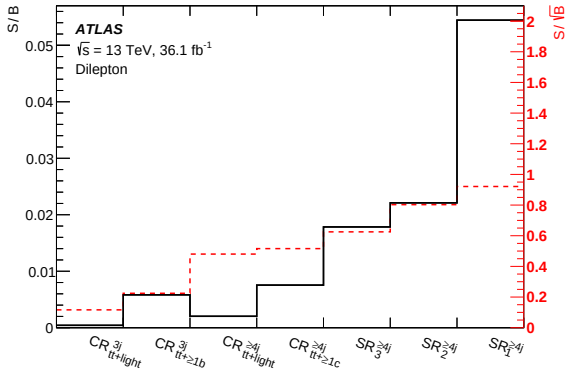
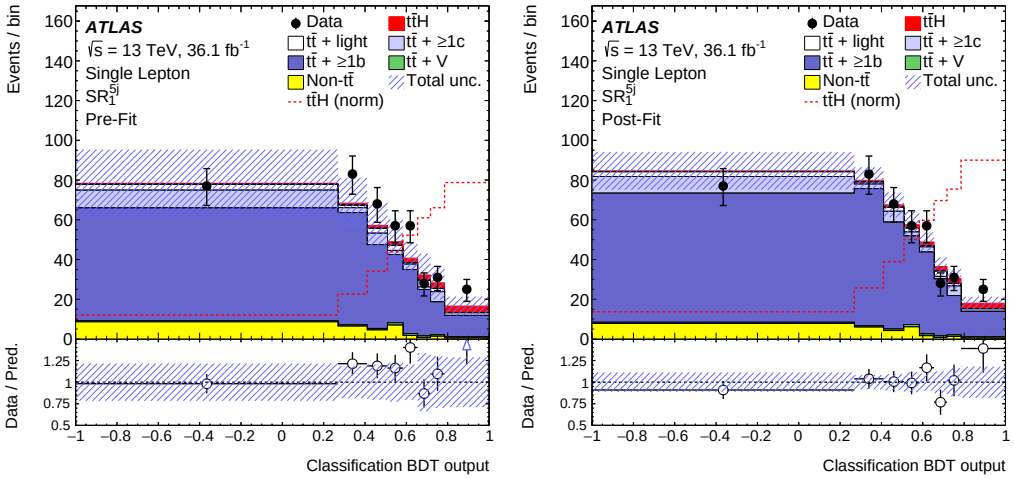
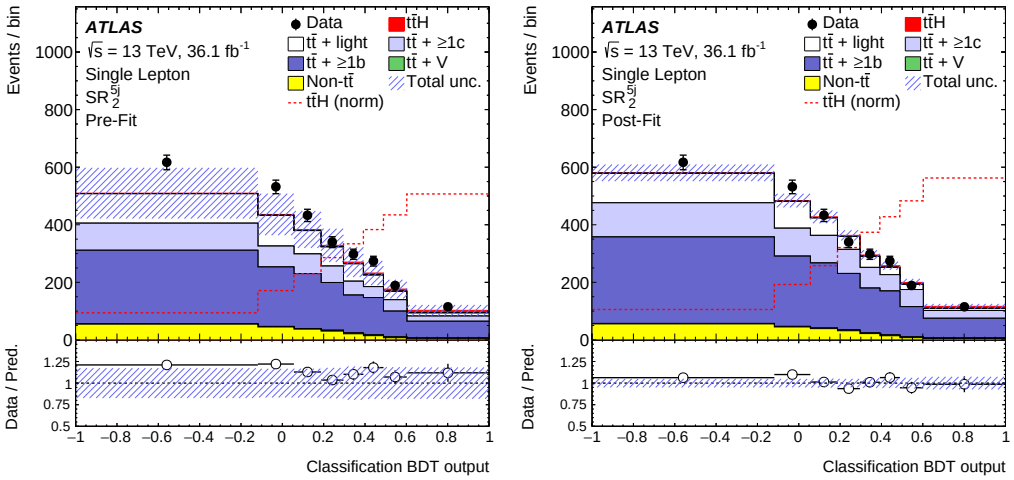


Figure A.3: Purity of the dileptonic signal and control regions [103]. The S/B ratio is shown in black on the left vertical axis and the $S/\sqrt{B}$ is shown in red on the right vertical axis.



(a) (b)



(c) (d)

Figure A.4: The distributions of the classification BDTs in the single-lepton signal regions with exactly five jets, pre-fit (a, c) and after the full combined fit (b, d) [103]. The $t\bar{t}H$ signal is shown in red stacked on top of the backgrounds where it is normalised to the SM cross-section pre-fit and to μ post-fit. The dashed red line shows the signal normalised to the total background prediction.

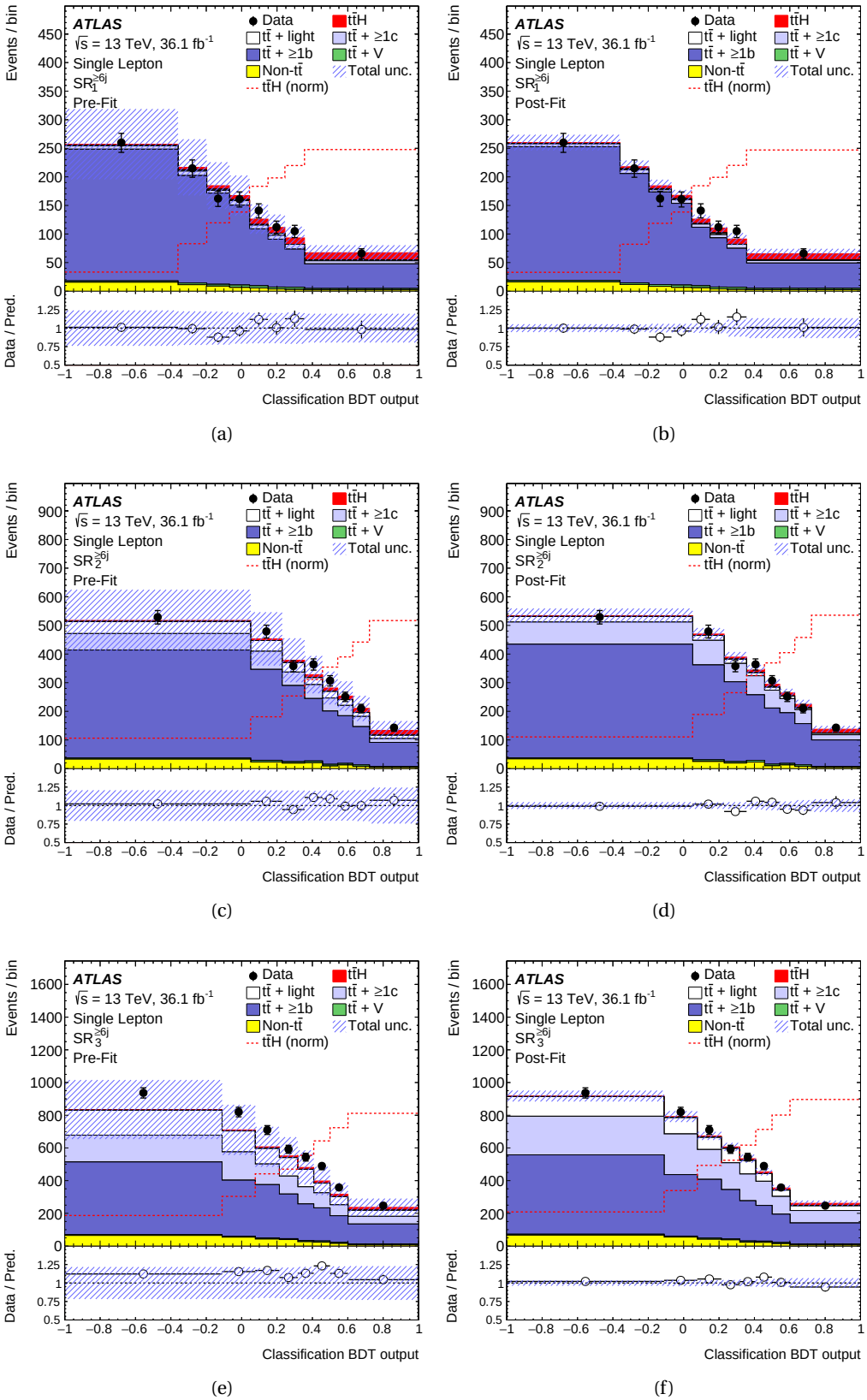


Figure A.5: The distributions of the classification BDTs in the single-lepton signal regions with at least six jets, pre-fit (a, c) and after the full combined fit (b, d) [103]. The $t\bar{t}H$ signal is shown in red stacked on top of the backgrounds where it is normalised to the SM cross-section pre-fit and to μ post-fit. The dashed red line shows the signal normalised to the total background prediction.

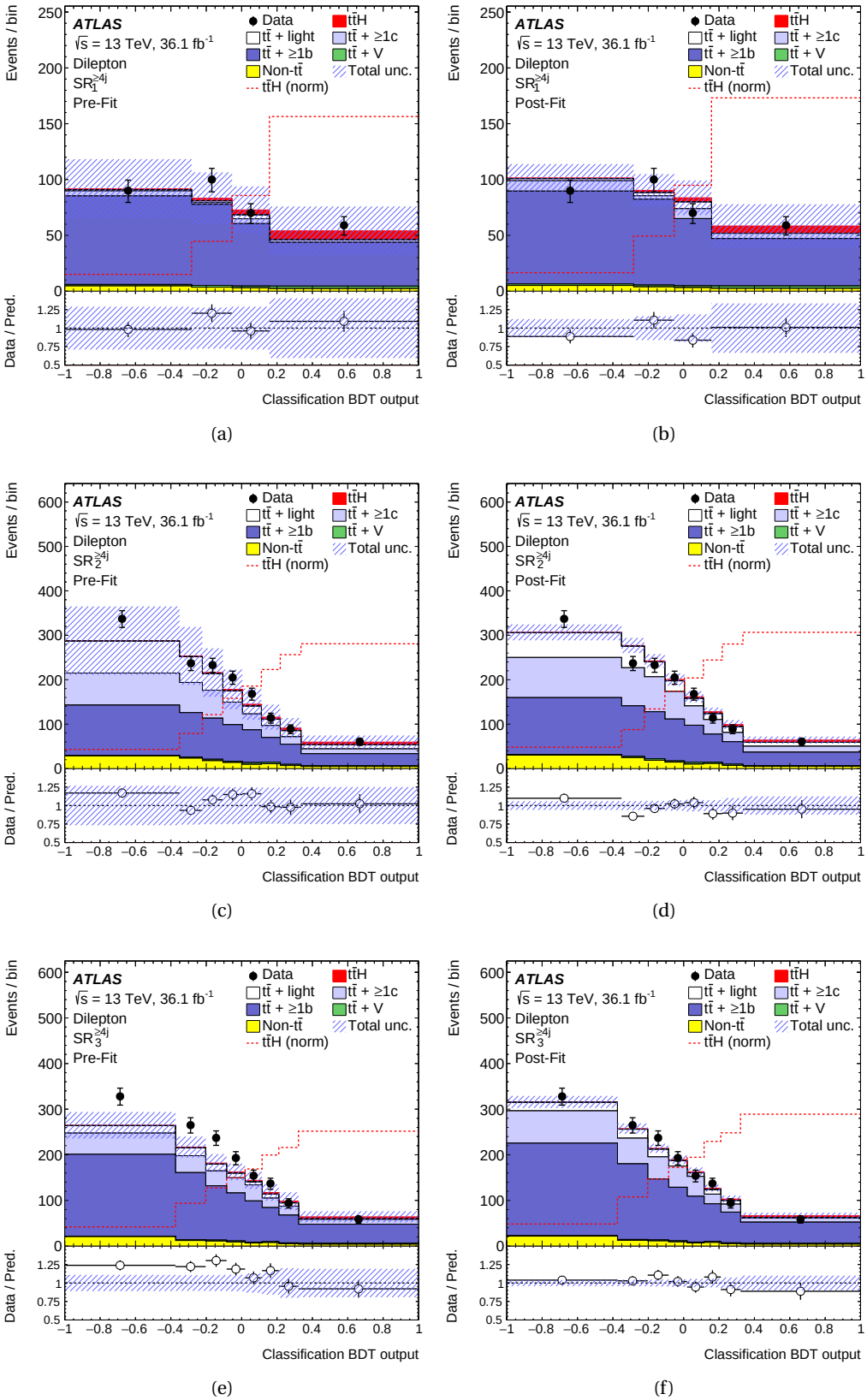
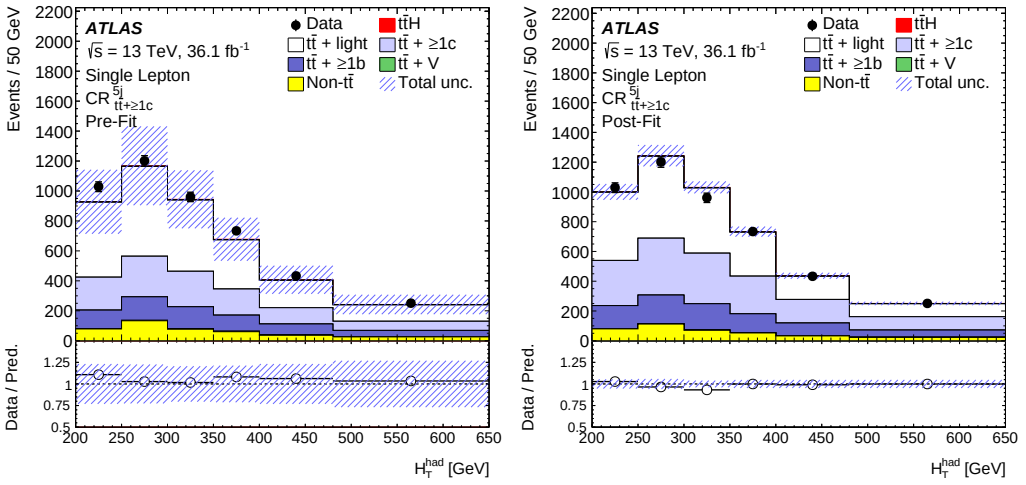


Figure A.6: The distributions of the classification BDTs in the dilepton signal regions with at least four jets, pre-fit (a, c) and after the full combined fit (b, d) [103]. The $t\bar{t}H$ signal is shown in red stacked on top of the backgrounds where it is normalised to the SM cross-section pre-fit and to μ post-fit. The dashed red line shows the signal normalised to the total background prediction.



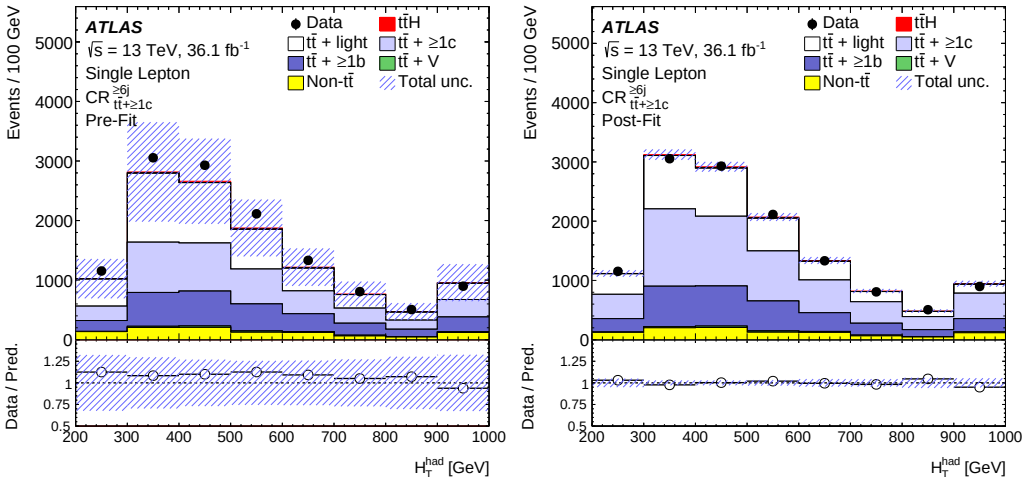


Figure A.7: The H_T^{had} distributions in the single-lepton $t\bar{t}+\geq 1c$ enriched control regions, pre-fit (a, c) and after the full combined fit (b, d) [103]. The $t\bar{t}H$ signal is shown in red and is normalised to the SM cross-section pre-fit and to μ post-fit.

GLOSSARY

4F four-flavour 35

5F five-flavour 35

AdaBoost Adaptive Boost 96

AF2 ATLAS FAST-II 38

ALICE A Large Ion Collider Experiment 20

ATLAS A Toroidal LHC ApparatuS 4

AUC Area Under the ROC Curve 98

BDT Boosted Decision Tree 43

BR branching ratio 73

BSM Beyond the Standard Model 14

C/A Cambridge/Aachen

CERN European Organization for Nuclear Research 19

CMS Compact Muon Solenoid 4

CP charge-parity 20

CSC Cathode Strip Chambers 32

DAQ data acquisition 21

ECAL electromagnetic calorimeter 28

EM electromagnetic 28

FCAL forward calorimeters 28

FSR final-state radiation 59

GSC global sequential calibration 51

HAD hadronic 29

HCAL hadronic calorimeter 28

HF heavy flavour 80

PS parton shower . . . 34

QCD Quantum Chromodynamics . . . 6

QED Quantum Electrodynamics . . . 6

ROC Receiver Operating Characteristics . . . 98

RoI Regions-of-Interest . . . 31

RPC Resistive Plate Chambers . . . 31

SCT Semi-Conductor Tracker . . . 25

SM Standard Model . . . 4

SPS Super Proton Synchrotron . . . 20

SR signal region . . . 82

SUSY supersymmetry . . . 20

TGC Thin Gap Chambers . . . 31

TRT Transition Radiation Tracker . . . 27

UE underlying event . . . 22

UV ultraviolet . . . 36

WP working point . . . 42

BIBLIOGRAPHY

[1] S. Glashow, *Partial symmetries of weak interactions*, Nucl. Phys. **22** (1961) 579–588. 4, 9

[2] S. Weinberg, *A model of leptons*, Phys. Rev. Lett. **19** no. 21, (1967). 4, 9

[3] A. Salam, *Gauge unification of fundamental forces*, Rev. Mod. Phys. **52** (1980) 525. 4, 9

[4] Particle Data Group, *The 2018 edition of the Review of Particle Physics*, Phys. Rev. D **98** no. 030001, (2018). 4, 5, 9, 12, 14, 15

[5] The ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** no. 1, (2012), arXiv:1207.7214. 4, 12, 14

[6] The CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716** no. 1, (2012), arXiv:1207.7235. 4, 12, 14

[7] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10** (1963) 531. 8

[8] M. Kobayashi and T. Maskawa, *CP-Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49** no. 2, (1973). 8

[9] F. Englert and R. Brout, *Broken symmetry and the mass of gauge vector mesons*, Phys. Rev. Lett. **13** no. 9, (1964) 321–323. 9, 12

[10] P. Higgs, *Broken symmetries and the masses of gauge bosons*, Phys. Rev. Lett. **13** no. 16, (1964) 508–509. 9, 12

[11] L. Álvarez-Gaumé and J. Ellis, *Eyes on a prize particle*, Nat. Phys. **7** no. 1, (2010). 10

[12] J. Goldstone, A. Salam, and S. Weinberg, *Broken symmetries*, Phys. Rev. **127** (1962) 965. 10

[13] The Economist, *Worth the wait*, 2012. https://www.economist.com/graphic-detail/2012/07/04/worth-the-wait. 12

[14] D. de Florian, C. Grojean, F. Maltoni, et al., *Handbook of LHC Higgs cross sections: 4. Deciphering the Nature of the Higgs Sector*, CERN2017–002-M (2016), arXiv:1610.07922. 13, 14, 15, 127, 151, 152, 153, 156

[15] The ATLAS and CMS Collaborations, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of*

the LHC pp collision data at $\sqrt{s} = 7$ *and 8 TeV*, J. High Energy Phys. **045** no. 08, (2016), arXiv:1606.02266. 14, 17

[16] The ATLAS Collaboration, *Observation of* $H \to b\bar{b}$ *decays and VH production with the ATLAS detector*, Phys. Lett. B **786** (2018) 59, arXiv:arXiv:1808.08238v2. 14

[17] The CMS Collaboration, *Observation of Higgs Boson decay to bottom quarks*, Phys. Rev. Lett. **121** no. 12, (2018), arXiv:1808.08242v2. 14

[18] The D0 Collaboration, *Observation of the Top Quark*, Phys. Rev. Lett. **74** (1995) 2632–2637, arXiv:hep-ex/9503003. 14

[19] The CDF Collaboration, *Observation of Top Quark Production in* $\bar{p}p$ *Collisions*, Phys. Rev. Lett. **74** (1995) 2626–2631, arXiv:hep-ex/9503002. 14

[20] M. Cacciari, M. Czakon, M. Mangano, A. Mitov, and P. Nason, *Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation*, Phys. Lett. B **710** no. 4-5, (2012), arXiv:1111.5869. 15, 79

[21] M. Czakon and A. Mitov, *NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels*, J. High Energy Phys. **54** (2012), arXiv:1207.0236. 15, 79

[22] M. Czakon and A. Mitov, *NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction*, J. High Energy Phys. **80** (2013), arXiv:1210.6832. 15, 79

[23] M. Czakon, P. Fiedler, and A. Mitov, *The total top quark pair production cross-section at hadron colliders through* $\mathcal{O}(\alpha_S^4)$, Phys. Rev. Lett. **110** (2013), arXiv:1303.6254. 15, 79

[24] The ATLAS Collaboration, *Search for a Higgs boson produced in association with a top-quark pair and decaying to* $b\bar{b}$ *in pp collisions at* $\sqrt{s} = 7$ *TeV using the ATLAS detector*, ATLAS-CONF-2012-135 (2012). https://cds.cern.ch/record/1478423. 17, 73, 132

[25] The ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into* $b\bar{b}$ *in pp collisions at* $\sqrt{s} = 8$ *TeV with the ATLAS detector*, Eur. Phys. J. C **075** no. 7, (2015), arXiv:1503.05066. 17, 73, 80, 114, 128, 132, 140

[26] The ATLAS Collaboration, *Search for the standard model Higgs boson decaying into* $b\bar{b}$ *produced in association with top quarks decaying hadronically in pp collisions at* $\sqrt{s} = 8$ *TeV with the ATLAS detector*, J. High Energy Phys. **160** no. 05, (2016), arXiv:1604.03812. 17, 73, 132

[27] The CMS Collaboration, *Search for the associated production of the Higgs boson with a top-quark pair*, J. High Energy Phys. **87** no. 09, (2014), arXiv:arXiv:1408.1682. 17

[28] *CERN: The accelerator complex*, 2018. https://home.cern/about/accelerators. 20

[29] *TE-EPC-LPC in LHC*, 2016. http://te-epc-lpc.web.cern.ch/te-epc-lpc/machines/lhc/general.stm. 20

[30] The ATLAS Collaboration, *ATLAS luminosity public results Run II*, 2019. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2. 21, 22

[31] The ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** no. 08, (2008). 22, 23, 24, 26, 27, 28, 29, 30, 31

[32] *CMS spherical coordinates, Universität Zürich*, 2017. https://wiki.physik.uzh.ch/cms/latex:example_spherical_coordinates. 23

[33] A. Ducourthial, M. Bomben, G. Calderini, et al., *Thin and edgeless sensors for ATLAS pixel detector upgrade*, JINST **12** no. 38, (2017), arXiv:1710.03557. 25, 28

[34] The ATLAS Collaboration, *ATLAS Insertable B-Layer Technical Design Report*, ATLAS TDR 19, Cern. 2010-013 (2010). 25, 26

[35] *ATLAS experiment public results - Event Displays from Collision Data*, 2015. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/EventDisplayRun2Collisions. 26

[36] The ATLAS collaboration, *Drift Time Measurement in the ATLAS Liquid Argon Electromagnetic Calorimeter using Cosmic Muons*, Eur. Phys. J. C **70** no. 3, (2010) 755–785. 29

[37] M. Livan and R. Wigmans, *Misconceptions about Calorimetry*, arXiv:1704.00661. 30

[38] S. Höche, *Introduction to parton-shower event generators*, SLAC-PUB 16160 (2014), arXiv:1411.4085. 35, 37

[39] R. D. Ball, V. Bertone, S. Carrazza, et al., *Parton distributions for the LHC run II*, J. High Energy Phys. **40** no. 04, (2015), arXiv:1410.8849. 36, 78, 130

[40] H.-L. Lai, M. Guzzi, J. Huston, et al., *New parton distributions for collider physics*, Phys. Rev. D **82** no. 7, (2010), arXiv:arXiv:1007.2241. 36, 84

[41] M. Guzzi, P. Nadolsky, E. Berger, et al., *CT10 parton distributions and other developments in the global QCD analysis*, SMU-HEP-10-11 (2010), arXiv:1101.0561v1. 36, 81, 130

[42] J. Gao, M. Guzzi, J. Huston, et al., *CT10 next-to-next-to-leading order global analysis of QCD*, Phys. Rev. D **89** no. 3, (2014) 033009. 36, 81, 130

[43] J. Alwall, R. Frederix, S. Frixione, et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, J. High Energy Phys. **79** no. 07, (2014), arXiv:1405.0301. 36, 78

[44] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with parton shower simulations: The POWHEG method*, J. High Energy Phys. no. 11, (2007), arXiv:0709.2092. 36, 79

[45] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: The POWHEG BOX*, J. High Energy Phys. **43** no. 06, (2010), arXiv:1002.2581. 36, 79

[46] T. Gleisberg, S. Höche, F. Krauss, et al., *Event generation with SHERPA 1.1*, J. High Energy Phys. **2** no. 7, (2009), arXiv:0811.4622. 36, 37, 81

[47] S. Frixione and B. R. Webber, *Matching NLO QCD computations and parton shower simulations*, J. High Energy Phys. no. 06, (2002), arXiv:hep-ph/0204244. 36

[48] P. Nason, *A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms*, J. High Energy Phys. no. 11, (2004), arXiv:hep-ph/0409146. 36

[49] F. Cascioli, P. Maierhöfer, and S. Pozzorini, *Scattering Amplitudes with Open Loops*, Phys. Rev. Lett. **108** (2012), arXiv:1111.5206v2. 36, 81

[50] T. Sjöstrand, S. Mrenna, and P. Skands, *PYTHIA 6.4 Physics and Manual*, J. High Energy Phys. no. 05, (2006), arXiv:hep-ph/0603175. 37, 82

[51] T. Sjöstrand, S. Mrenna, and P. Skands, *A brief introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** no. 11, (2008) 852–867, arXiv:0710.3820. 37, 78

[52] G. Corcella, I. Knowles, G. Marchesini, et al., *HERWIG 6: an event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)*, J. High Energy Phys. no. 01, (2001), arXiv:hep-ph/0011363v1. 37

[53] M. Bahr, S. Gieseke, M. A. Gigg, et al., *Herwig++ Physics and Manual*, Eur. Phys. J. C **58** no. 4, (2008) 639–707, arXiv:0803.0883. 37, 84

[54] T. Sjöstrand and P. Skands, *Transverse-Momentum-Ordered Showers and Interleaved Multiple Interactions*, Eur. Phys. J. C **39** (2005) 129–154, arXiv:hep-ph/0408302. 37

[55] B. Andersson, *Recent Developments in the Lund Model*, arXiv:hep-ph/0212122v1. 37

[56] A. Kupčo, *Cluster Hadronization in Herwig 5.9*, arXiv:hep-ph/9906412. 37

[57] The ATLAS Collaboration, *Measurement of the underlying event in jet events from 7 TeV proton-proton collisions with the ATLAS detector*, Eur. Phys. J. C **74** no. 2965, (2014), arXiv:1406.0392. 37

[58] The GEANT4 Collaboration, *GEANT4: A simulation toolkit*, Nucl. Instruments Methods Phys. A **506** no. 3, (2003) 250–303. 38

[59] W. Lukas, *Fast simulation for ATLAS: Atlfast-II and ISF*, J. Phys. Conf. Ser. **396** no. 022031, (2012). 38

[60] The ATLAS Collaboration, *The ATLAS calorimeter simulation FastCaloSim*, J. Phys. Conf. Ser. **331** no. 032053, (2011). 38

[61] The ATLAS Collaboration, *ATLAS Computing : technical design report*, ATLAS-TDR-17 (2005). https://cds.cern.ch/record/837738. 40

[62] T. Cornelissen, M. Elsing, I. Gavrilenko, et al., *Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)*, ATL-SOFT-PUB-2007-007 (2007). https://cds.cern.ch/record/1020106?ln=en. 40, 41

[63] The ATLAS Collaboration, *The Optimization of ATLAS Track Reconstruction in Dense Environments*, ATL-PHYS-PUB-2015-006 (2015). https://cds.cern.ch/record/2002609. 40

[64] The ATLAS Collaboration, *Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton-proton collisions at the LHC*, Eur. Phys. J. C **77** no. 5, (2017), arXiv:1611.10235. 40

[65] The ATLAS Collaboration, *Performance of primary vertex reconstruction in proton-proton collisions at* $\sqrt{s} = 7$ *TeV in the ATLAS experiment*, ATLAS-CONF-2010-069 (2010). http://cds.cern.ch/record/1281344. 41

[66] The ATLAS Collaboration, *Electron reconstruction and identification efficiency measurements with the atlas detector using the 2011 LHC proton-proton collision data*, Eur. Phys. J. C **74** no. 7, (2014), arXiv:1404.2240. 41

[67] The ATLAS Collaboration, *Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data*, Eur. Phys. J. C **74** no. 10, (2014), arXiv:1407.5063. 41

[68] The ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using 2015 LHC proton-proton collision data*, ATLAS-CONF-2016-024 (2015). https://cds.cern.ch/record/2157687. 41, 42, 75, 76, 78, 127

[69] The ATLAS Collaboration, *Measurements of the photon identification efficiency with the ATLAS detector using 4.9* fb^{-1} *of pp collision data collected in 2011*, ATLAS-CONF-2012-123 (2012). https://cds.cern.ch/record/1473426. 42

[70] The ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton-proton collision data at* $\sqrt{s} = 13$ *TeV*, Eur. Phys. J. C **76** no. 5, (2016), arXiv:1603.05598. 42, 75, 76, 78, 127

[71] The ATLAS Collaboration, *Reconstruction, Energy Calibration, and Identification of Hadronically Decaying Tau Leptons in the ATLAS Experiment for Run-2 of the LHC*, ATL-PHYS-PUB-2015-045 (2015). https://cds.cern.ch/record/2064383. 43, 76, 78

[72] The ATLAS Collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector in the first proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, ATL-PHYS-PUB-2015-027 (2015). https://cds.cern.ch/record/2037904. 43

[73] The ATLAS collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, Eur. Phys. J. C **77** (2017) 490, arXiv:1603.02934v3. 44, 45, 46

[74] G. Sterman and S. Weinberg, *Jets from Quantum Chromodynamics*, Phys. Rev. Lett. **39** (1977) 1436. 46

[75] S. Catani, Y. Dokshitzer, M. Seymour, and B. Webber, *Longitudinally-invariant* $k_\perp$*-clustering algorithms for hadron-hadron collisions*, Nucl. Physics, Sect. B **406** no. 1-2, (1993) 187–224. 47

[76] S. Ellis and D. Soper, *Successive combination jet algorithm for hadron collisions*, Phys. Rev. D **48** no. 7, (1993) 3160–3166, arXiv:hep-ph/9305266. 47

[77] M. Cacciari, G. Salam, and G. Soyez, *The anti-*k_t *jet clustering algorithm*, J. High Energy Physics, **04** (2008) 12, arXiv:0802.1189. 47, 48

[78] Y. Dokshitzer, G. Leder, S. Moretti, and B. Webber, *Better Jet Clustering Algorithms*, J. High Energy Phys. **1997** no. 08, (1997), arXiv:hep-ph/9707323. 48

[79] M. Wobisch and T. Wengler, *Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering*, in *Proc. Work. Monte Carlo Gener. HERA Phys.* 1999. arXiv:hep-ph/9907280. 48

[80] L. Asquith, B. Brelier, J. Butterworth, et al., *Performance of Jet Algorithms in the ATLAS Detector*, ATL-PHYS-INT-2010-129 (2010). https://cds.cern.ch/record/1311867? 48

[81] M. Cacciari, G. Salam, and G. Soyez, *FastJet user manual*, Eur. Phys. J. C **72** (2012) 1896, arXiv:1111.6097. 49, 76

[82] The ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Phys. Rev. D **96** no. 072002, (2017), arXiv:1703.09665. 49, 51, 52, 123

[83] The ATLAS Collaboration, *Pile-up subtraction and suppression for jets in ATLAS*, ATLAS-CONF-2013-083 (2013). https://cds.cern.ch/record/1570994. 50

[84] M. Cacciari and G. Salam, *Pileup subtraction using jet areas*, Phys. Lett. B **659** no. 1-2, (2008) 119–126, arXiv:0707.1378. 50

[85] The ATLAS Collaboration, *Jet global sequential corrections with the ATLAS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV*, ATLAS-CONF-2015-002 (2015). http://cdsweb.cern.ch/record/2001682. 51

[86] The ATLAS Collaboration, *Selection of jets produced in 13 TeV protonproton collisions with the ATLAS detector*, ATLAS-CONF-2015-029 (2015). https://cds.cern.ch/record/2037702. 52

[87] The ATLAS Collaboration, *Tagging and suppression of pileup jets with the ATLAS detector*, ATLAS-CONF-2014-018 (2014). https://cds.cern.ch/record/1700870. 53

[88] The ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector*, Eur. Phys. J. C **76** no. 11, (2016) 581, arXiv:1510.03823. 53

[89] The ATLAS Collaboration, *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11** no. April, (2015), arXiv:1512.01094. 54, 124

[90] The ATLAS Collaboration, *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, ATL-PHYS-PUB-2016-012 (2016). https://cds.cern.ch/record/2160731. 54, 55

[91] The ATLAS Collaboration, *Commissioning of the ATLAS b-tagging algorithms using $t\bar{t}$ events in early Run-2 data*, ATL-PHYS-PUB-2015-039 (2015). https://cds.cern.ch/record/2047871. 54

[92] G. Piacquadio and C. Weiser, *A new inclusive secondary vertex algorithm for b-jet tagging in ATLAS*, J. Phys. Conf. Ser. **119** no. 3, (2008) 032032. 55

[93] J. Thaler and K. Van Tilburg, *Identifying boosted objects with N-subjettiness*, J. High Energy Phys. **15** no. 03, (2011), arXiv:1011.2268. 56

[94] J. Thaler and K. Van Tilburg, *Maximizing boosted top identification by minimizing N-subjettiness*, J. High Energy Phys. **93** no. 02, (2012), arXiv:1108.2701. 56

[95] J. Butterworth, B. Cox, and J. Forshaw, *WW scattering at the CERN LHC*, Phys. Rev. D **65** no. 9, (2002) 096014, arXiv:hep-ph/0201098 [hep-ph]. 57

[96] The ATLAS Collaboration, *Boosted hadronic top identification at ATLAS for early 13 TeV data*, ATL-PHYS-PUB-2015-053 (2015). https://cds.cern.ch/record/2116351. 58

[97] D. Krohn, J. Thaler, and L.-T. Wang, *Jet Trimming*, J. High Energy Phys. **84** no. 02, (2010), arXiv:0912.1342. 59

[98] The ATLAS Collaboration, *Identification of Boosted , Hadronically-Decaying W and Z Bosons in $\sqrt{s}$ = 13 TeV Monte Carlo Simulations for ATLAS*, ATLAS-PHYS-PUB-2015-033 (2015). 59

[99] The ATLAS Collaboration, *Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s}$ = 7 TeV using the ATLAS detector*, J. High Energy Phys. **2013** no. 9, (2013), arXiv:1306.4945. 59, 60

[100] The ATLAS Collaboration, *Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s}$=7 TeV using the ATLAS detector*, J. High Energy Phys. **076** no. 09, (2013), arXiv:1306.4945. 60

[101] B. Nachman, P. Nef, A. Schwartzman, M. Swiatlowski, and C. Wanotayaroj, *Jets from Jets: Re-clustering as a tool for large radius jet reconstruction and grooming at the LHC*, J. High Energy Phys. **75** no. 02, (2015), arXiv:1407.2922 [hep-ph]. 60

[102] The ATLAS Collaboration, *Jet reclustering and close-by effects in ATLAS Run 2*, ATLAS-CONF-2017-062 (2017). https://cds.cern.ch/record/2275649. 61

[103] The ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Phys. Rev. D **97** no. 7, (2018), arXiv:1712.08895v2. 73, 81, 93, 94, 95, 141, 146, 147, 148, 150, 160, 161, 162, 163, 164

[104] The ATLAS Collaboration, *Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton-proton collision data at $\sqrt{s}$ = 13 TeV collected with the ATLAS experiment*, ATLAS-CONF-2018-31 (2018). https://cds.cern.ch/record/2625365. 73

[105] The ATLAS Collaboration, *Evidence for the associated production of the Higgs boson and top quark pair with the ATLAS detector*, Phys. Rev. D **97** no. 7, (2018), arXiv:1712.08891. 78, 151, 153, 154

[106] The ATLAS Collaboration, *ATLAS Standard Model Physics public results*, 2019. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/StandardModelPublicResults. 80

[107] The ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021 (2014). https://cds.cern.ch/record/1974411. 78

[108] S. Heinemeyer, C. Mariotti, G. Passarino, and R. Tanaka, *Handbook of LHC Higgs cross sections: 3. Higgs Properties*, arXiv:arXiv:1307.1347. 78

[109] The ATLAS Collaboration, *Studies on top-quark Monte Carlo modelling for Top2016*, ATL-PHYS-PUB-2016-020 (2016). https://cds.cern.ch/record/2216168. 79, 128, 129

[110] M. Czakon and A. Mitov, *Top++: A program for the calculation of the top-pair cross-section at hadron colliders*, Comput. Phys. Commun. **185** no. 11, (2014). 79, 128

[111] F. Cascioli, P. Maierhöfer, N. Moretti, S. Pozzorini, and F. Siegert, *NLO matching for $t\bar{t}-b\bar{b}$ production with massive b-quarks*, Phys. Lett. B **734** (2014) 210–214, arXiv:1309.5912. 81

[112] P. Skands, *Tuning Monte Carlo generators: The Perugia tunes*, Phys. Rev. D **82** no. 7, (2010) 1–46, arXiv:1005.3457. 82

[113] The ATLAS Collaboration, *Estimation of non-prompt and fake lepton backgrounds in final states with top quarks produced in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, ATLAS-CONF-2014-058 (2014). https://cds.cern.ch/record/1951336. 84

[114] A. Hoecker, P. Speckmayer, J. Stelzer, et al., *TMVA 4 - Toolkit for Multivariate Data Analysis with ROOT*, CERN-OPEN-2007-007 (2007), arXiv:physics/0703039/. 96, 97, 110, 114

[115] The ATLAS Collaboration, *Search for flavour-changing neutral current top quark decays $t \to Hq$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, J. High Energy Phys. **61** no. 12, (2015), arXiv:1509.06047. 113

[116] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011), arXiv:1007.1727v3. 115, 116, 118, 119

[117] The ATLAS Collaboration, The CMS Collaboration, and The LHC Higgs Combination Group, *Procedure for the LHC Higgs boson search combination in Summer 2011*, ATL-PHYS-PUB-2011-011 (2011). http://cds.cern.ch/record/1379837/files/NOTE2011_005.pdf?version=1. 115

[118] J. Neyman and E. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, Philos. Trans. R. Soc. London. Ser. A, Contain. Pap. a Math. or Phys. Character **231** (1933) 289–337. 116

[119] T. Junk, *Confidence level computation for combining searches with small statistics*, Nucl. Instruments Methods Phys. Res. Sect. A **434** no. 2, (1999) 435–443, arXiv:hep-ex/9902006. 119

[120] A. Read, *Presentation of search results : the CLs technique*, J. Phys. G **28** no. 10, (2002). 119

[121] S. van der Meer, *Calibration of the effective beam height in the ISR*, ISR-PO/68-31, KEK68-64 (1968). 123

[122] The ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **76** (2016) 653, arXiv:1608.03953. 123

[123] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Parton distributions for the LHC*, Eur. Phys. J. C **63** no. 2, (2009), arXiv:0901.0002. 130

[124] W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, in *CHEP03 Proc.* 2003. arXiv:physics/0306116. 133

[125] The ROOT Math Library Team, "Minuit2 Minimization Package." http://project-mathlibs.web.cern.ch/project-mathlibs/sw/Minuit2/html/index.html. 133

[126] The ATLAS Collaboration, *Measurements of Higgs boson properties in the diphoton decay channel with $36\,fb^{-1}$ of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **98** no. 5, (2018), arXiv:1802.04146. 151, 153

[127] The ATLAS Collaboration, *Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector*, J. High Energy Phys. **095** no. 03, (2018), arXiv:1712.02304. 151, 153

[128] The ATLAS Collaboration, *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*, Phys. Lett. B **784** (2018) 173–191, arXiv:1806.00425. 153, 155, 156

[129] The ATLAS Collaboration, *Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment*, Eur. Phys. J. C **76** no. 6, (2016), arXiv:1507.04548v3. 153

[130] The ATLAS Collaboration, *Electron and photon reconstruction and performance in ATLAS using a dynamical, topological cell clustering-based approach*, ATL-PHYS-PUB-2017-022 (2017). https://cds.cern.ch/record/2298955. 153